UNIVERSITY OF HELSINKI
FACULTY OF MEDICINE

INSTRUMENTS FOR IMAGE QUALITY ESTIMATION

TONI VIRTANEN

Department of Psychology and Logopedics
Faculty of Medicine
University of Helsinki

INSTRUMENTS FOR IMAGE QUALITY ESTIMATION

KUVANLAATUKOKEMUKSEN ARVIONNIN INSTRUMENTIT

Toni Virtanen

DOCTORAL DISSERTATION

Doctoral dissertation, to be presented for public discussion with the permission of the Faculty of Medicine of the University of Helsinki, in Auditorium PII, Porthania building, on the 13th of August, 2020, at 12 o'clock.

Helsinki 2020

Supervisors

Docent Jukka Häkkinen, PhD
Department of Psychology and Logopedics
Faculty of Medicine
University of Helsinki, Finland

Professor Emeritus Göte Nyman, PhD
Department of Psychology and Logopedics
Faculty of Medicine
University of Helsinki, Finland

Reviewers

Reader Sophie Triantaphillidou, PhD
Computer Science Department
Faculty of Science and Technology
University of Westminster, United Kingdom

Associate Professor Damon Chandler, PhD
Department of Electrical and Electronic Engineering
Faculty of Engineering
Shizouka University, Japan

Opponent

Professor Marius Pedersen, PhD
Department of Computer Science
Faculty of Information Technology and Electrical Engineering
Norwegian University of Science and Technology, Norway

ISBN 978-951-51-6361-5 (pbk.)
ISBN 978-951-51-6362-2 (PDF)

Unigrafia
Helsinki 2020

ABSTRACT

This dissertation describes the instruments available for image quality evaluation, develops new methods for subjective image quality evaluation and provides image and video databases for the assessment and development of image quality assessment (IQA) algorithms. The contributions of the thesis are based on six original publications. The first publication introduced the VQone toolbox for subjective image quality evaluation. It created a platform for free-form experimentation with standardized image quality methods and was the foundation for later studies. The second publication focused on the dilemma of reference in subjective experiments by proposing a new method for image quality evaluation: the absolute category rating with dynamic reference (ACR-DR).

The third publication presented a database (CID2013) in which 480 images were evaluated by 188 observers using the ACR-DR method proposed in the prior publication. Providing databases of image files along with their quality ratings is essential in the field of IQA algorithm development.

The fourth publication introduced a video database (CVD2014) based on having 210 observers rate 234 video clips. The temporal aspect of the stimuli creates peculiar artifacts and degradations, as well as challenges to experimental design and video quality assessment (VQA) algorithms. When the CID2013 and CVD2014 databases were published, most state-of-the-art I/VQAs had been trained on and tested against databases created by degrading an original image or video with a single distortion at a time. The novel aspect of CID2013 and CVD2014 was that they consisted of multiple concurrent distortions.

To facilitate communication and understanding among professionals in various fields of image quality as well as among non-professionals, an attribute lexicon of image quality, the image quality wheel, was presented in the fifth publication of this thesis. Reference wheels and terminology lexicons have a long tradition in sensory evaluation contexts, such as taste experience studies, where they are used to facilitate communication among interested stakeholders; however, such an approach has not been common in visual experience domains, especially in studies on image quality.

The sixth publication examined how the free descriptions given by the observers influenced the ratings of the images. Understanding how various elements, such as perceived sharpness and naturalness, affect subjective image quality can help to understand the decision-making processes behind image quality evaluation. Knowing the impact of each preferential attribute can then be used for I/VQA algorithm development; certain I/VQA algorithms already incorporate low-level human visual system (HVS) models in their algorithms.

TIIVISTELMÄ

Väitöskirja tarkastelee kuvanlaadun arviointiin käytettävissä olevia instrumentteja, kehittää uusia menetelmiä subjektiiviseen kuvanlaadun arviointiin sekä tarjoaa kuva- ja videotietokantoja kuvanlaadun arviointialgoritmien (IQA) testaamiseen ja kehittämiseen. Tutkielma on jaettu kuuteen alkuperäiseen julkaisuun.

Ensimmäisessä julkaisussa kehitettiin Matlab VQone -ohjelmisto subjektiiviselle kuvanlaadun arvioinnille tutkijoiden vapaaseen käyttöön. Se antoi mahdollisuuden testata standardoituja kuvanlaadun arviointiin kehitettyjä menetelmiä ja kehittää niiden pohjalta myös uusia menetelmiä luoden perustan myöhemmille tutkimuksille. Toisessa julkaisussa kehitettiin uusi subjektiivinen kuvanlaadun arviointimenetelmä: ”absolute category rating with dynamic reference” (ACR-DR). Menetelmä hyödyntää sarjallista kuvien esitystapaa, jolla muodostettiin arvioijille mielikuva kuvien laatuvaihtelusta ennen varsinaista laatuarviointia. Menetelmän todettiin vähentävän tulosten hajontaa ja erottelevan pienempiä kuvanlaatueroja.

Kolmannessa julkaisussa kuvaillaan tietokanta, jossa on 188 henkilön 480 kuvasta ACR-DR-menetelmällä tekemät laatuarviot ja niihin liittyvät kuvatiedostot.

Neljännessä julkaisussa esitellään tietokanta, jossa on 210 henkilön 234 videoleikkeestä tekemät laatuarviot ja niihin liittyvät videotiedostot. Ajallisen ulottuvuuden vuoksi videoärsykkeiden virheet ovat erilaisia kuin kuvissa, mikä tuo omat haasteensa subjektiivisen kuvanlaadun kokeiden suunnitteluun. Se on myös haasteellista videoiden laatua arvioiville algoritmeille (VQA). Aikaisempien kuva- ja videotietokantojen sisältö on luotu vääristämällä hyvälaatuista alkuperäistä ärsykettä yksi vääristymä kerrallaan. Tämä on tehty esimerkiksi kuvaa tai videota asteittain sumentamalla. Tässä väitöskirjassa esitetyt tietokannat poikkeavat aikaisemmista, sillä ne on kuvattu eri kameroilla ilman jälkikäteen tehtyä kuvanmuokkausta. Niinpä ne koostuivat kuvista ja videoista, jotka sisältävät useita samanaikaisia vääristymistä.

Viidennessä julkaisussa esitellään kuvanlaatuympyrä (image quality wheel). Se on kuvanlaadun käsitteiden sanasto, joka on kerätty analysoimalla 146 henkilön tuottamat 39 415 kuvanlaadun sanallista kuvausta. Sanastoilla on pitkät perinteet aistinvaraisen arvioinnin tutkimusperinteessä, mutta niitä ei ole aikaisemmin kehitetty visuaaliselle kuvanlaadulle.

Kuudennessa tutkimuksessa tutkittiin, kuinka arvioitsijoiden antamat käsitteet vaikuttavat kuvien laadun arviointiin. Esimerkiksi kuvien arvioitu terävyys tai luonnollisuus auttaa ymmärtämään laadun arvioinnin taustalla olevia päätöksentekoprosesseja. Tietoa voidaan käyttää esimerkiksi kuvan- ja videonlaadun arviointialgoritmien (I/VQA) kehitystyössä.

ACKNOWLEDGEMENTS

Most importantly I want to give sincere thanks for my honorable opponent Professor Marius Pedersen, as well as the two pre-examiners of this doctoral dissertation, Associate Professor Damon Chandler and Reader Sophie Triantaphillidou. Receiving constructive criticism from professionals who I look up to is a joy. I will also want to give my thanks to the Custos, Professor Kimmo Alho as the representative of the University of Helsinki, Faculty of Medicine.

I would not be writing this without my mentors Docent Jukka Häkinen and Emeritus Professor Göte Nyman. I owe it to Jukka, who patiently guided me through the majority of this process. Jukka's keen insights and comments gave it focus, that I admittetly seem to sometimes lack and tend to get distracted again and again about some new project or idea I might have. Göte Nyman was the principal supervisor of this dissertation until his retirement. I feel priviledged to have been able to work with him. Your thoughs on humanity, technological progress, academic sincerity and life long curiosity towards learning new things is still inspiring.

I will want to give my special thank to two of my colleagues in particular: Mikko Nuutinen, you really helped me raise the level of this disseration by introducing me to the field of computational image quality assessment and algorithmic thinking. Jenni Radun, you taught me how to conduct qualitative analysis, the Interpretation-based Quality (IBQ) in particular, and how word frequencies could be statistically analysed and combined with numerical ratings. I like to think this dissertation is a synthesis of things I've learned from both of you.

I wish to thank all my co-authors of the original communications, Pirkko Oittinen, Mikko Vaahteranoksa, Tero Vuori, Terhi Mustonen, Tuomas Leisti and Olli Rummukainen. It has been a priviledge to work with you. I would also like to thank the anonymous reviewers of the articles sent for peer review under this dissertation project.

I will also want to give my thanks to my colleagues at our research group that I have had the pleasure to work with. Tuomas Leisti, Terhi Mustonen, Olli Rummukainen, Jari Takatalo, Jyrki Kaistinen, Oskari Salmi, Timo Säämänen, Paul Lindroos, Perttu Pöyhönen, Anna Toni, Eero-Matti Gummerus (nee Koivisto), Esa Nygren (nee Anttonen), Dana Vainikka (nee Kostik), Jaakko Airaksinen, Sini Hämäläinen (nee Jakonen), Eero Iso-Kokkila, Jaakko Tähkä, Milla Huuskanen, Suvi Hoffman (nee Holm), Hanna Weckman, Jussi Hakala, Hannu Alén. I counted that a total of 651 anonymous opbservers participated to the experiments in this dissertation and wish to thank them all. Many of my colleagues mentioned above also aided me with the experiments during this process and without them I would still be in the lab overseeing the experiments.

This work would also not have been made possible without our collaboration with industry partners in Nokia and Microsoft, and I want to specially thank Tero Vuori, Mikko Vaahteranoksa, Jean-Luc Olives, Ari Sirén and Joni Oja for all those years. It might even seem a bit backwards, but without our industry partners we would probably never have started such a close collaboration with the Visual Media research group at the Aalto University, led then by Professor Pirkko Oittinen. This disseration would probably not exist without that inspiring interdiciplinary academic-industry environment that were created back then.

I will also want to thank all my friends and colleagues at the Finnish Defence Research Agency. You've given me your support as I've lived through the ups and downs of the final stretches of this project.

This dissertation was funded by the Graduate School in User-Centered Information Technology (UCIT) and the HPY Research Foundation. Additional funding and support came through industry partners, Nokia and Microsoft, as many of the experiments and stimuli were related to the various projects we worked on during the years.

Thanks also to all my friends for reminding me that life is not just about work. Last but not least, I want to give my sincere thanks to my family who gave me a safe and loving environment to grow. To my spouse Ulla, thank you for your unconditional support and understanding.

Helsinki, 2020
Toni Virtanen

This disseration is dedicated to the loving memory of my father Unto Virtanen who passed away just a few months before its publication – I miss you.

CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following publications:

I. Nuutinen, M., Virtanen, T., Rummukainen, O., & Häkkinen, J. (2016). VQone MATLAB toolbox: A graphical experiment builder for image and video quality evaluations. *Behavior Research Methods*, *48*(1).

II. Nuutinen, M., Virtanen, T., Leisti, T., Mustonen, T., Radun, J., & Häkkinen, J. (2016). A new method for evaluating the subjective image quality of photographs: dynamic reference. *Multimedia Tools and Applications*, *75*(4).

III. Virtanen, T., Nuutinen, M., Vaahteranoksa, M., Oittinen, P., & Häkkinen, J. (2015). CID2013: a database for evaluating no-reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, *24*(1).

IV. Nuutinen, M., Virtanen, T., Vaahteranoksa, M., Vuori, T., Oittinen, P., & Häkkinen, J., (2016). CVD2014 - a database for evaluating no-reference video quality assessment algorithms. *IEEE Transactions on Image Processing*, *25*(7).

V. Virtanen, T., Nuutinen, M., & Häkkinen, J. (2019). Image quality wheel. *Journal of Electronic Imaging*, *28*(1).

VI. Virtanen, T., Nuutinen, M., & Häkkinen, J. (2020). Underlying elements of image quality assessment: Preference and terminology for communicating image quality characteristics. *Psychology of Aesthetic, Creativity, and the Arts. Advance online publication (2020, April 9).*

Publication I.

The author supervised the MatLab toolbox development and was closely involved in devising many of its key functions such as the new method, i.e., Dynamic Reference Absolute Category Rating (ACR-DR), the random starting points for the sliders, the free-form experimental build panel and the response plot visualization for the participant. The base of the program was written by Olli Rummukainen and finished by Mikko Nuutinen, who was also the first author of the publication. The author of the dissertation was the second author of the publication.

Publication II.

The author was one of the originators of the idea behind the newly presented ACR-DR method for evaluating images with a dynamic reference. The author contributed to the design and implementation of the validation experiments. The author of the dissertation was the second author of the publication.

Publication III.

The author designed and supervised all the subjective experiments conducted by undergraduate research aides. The author conducted necessary statistical tests for the publication and was the main author of the publication.

Publication IV.

The author designed and supervised subjective experiments 1 to 6 and contributed to the design and implementation of the re-alignment experiment. The author contributed to the statistical testing for the publication based on publication III. The author of the dissertation was the second author of the publication.

Publication V.

The author designed and supervised all the subjective experiments conducted by undergraduate research aides. The author conducted all the text analyses, implemented natural language processing methods for the free descriptions, and was the main author of the publication.

Publication VI.

The author designed and supervised all the subjective experiments conducted by undergraduate research aides. The author conducted all statistical tests for the publication and was the main author of the publication.

ACRONYMS

***nesses**	Preferential attributes such as sharpness and colorfulness
2-AFC	Two-alternative forced choice, a forced choice type of paired comparison task
3A	Auto focus, auto exposure and auto white balance
4K	Refers to a display resolution of approximately 4000 pixels
8K	Refers to a display resolution of approximately 8000 pixels
A4	Size A4 paper, 210 mm × 297 mm (8.27 in × 11.7 in)
ACR	Absolute category rating
ACR-DR	Absolute category rating with dynamic reference
ACR-HR	Absolute category rating with hidden reference
AE	Auto exposure
AF	Auto focus
AWB	Auto White Balance
AVC HD Database	High-definition H.264/AVC video database
BIB	Balanced incomplete block design, a method to balance the comparison combinations of the stimuli to minimize the experimental time
BID	Blurred image database
BIQI	Blind image quality index
BLIINDS-II	BLIind non-distortion specific VQA algorithm
BRISQUE	Blind/referenceless image spatial quality evaluator
cd/m^2	Candela per square meter
CID2013	Camera image database 2013
CIELAB	Colorspace created by the International Comission of Lighting.
CORNIA	Codebook representation for no-reference image assessment
CPBD	Cumulative probability of blur detection iqa algorithm
CPIQ	Camera phone image quality initiative working group
CRT	Cathode ray tube display
CSF	Contrast sensitivity function is a measure of the ability to discern between luminances of different levels in a static image.
CSIQ	Categorical subjective image quality database
CVD2014	Camera Video Database 2014

DCR	Degradation category rating
DESIQUE	DErivative Statistics-based QUality Evaluator an NR-IQA algorithm
df	Degrees of freedom
DIIVINE	Distortion identification-based image verity and integrity evaluation index, an NR-IQA algorithm
DMOS	Differential mean opinion score
DSCQS	Double stimulus continuous quality scale
DSIS	Double stimulus impairment scale
DSLR	Digital single lens reflex camera
DVC	Digital video camera
ECVQ	CIF Video Quality database by University of Osijek
EFPL-PoliMi	Video database by École Polytechnique Fédérale de Lausanne and Politecnico di Milano
EVVQ	VGA video quality database by University of Osijek
F.A.C.T.	Functional acuity contrast test
FISH	Fast image sharpness, an IQA algorithm focusing on estimating sharpness
FISH_bb	Fast image sharpness, a local-block based variation of the FISH algorithm
fps	Frames per second
FR	Full-reference
FR-IQA	Full-reference image quality assessment algorithm
GUI	Graphical user interface
HVS	Human visual system
IEEE	Institute of Electrical and Electronics Engineers
I/VQA	Image and/or video quality assessment algorithm
I3A	International imaging industry association
IBM	International Business Machines Corporation
IBQ	Interpretation-based quality
ICC	International Color Consortium
IQA	Image quality assessment algorithm
IRCCyN	Institut de Recherche en Communications et Cybernétique de Nantes
ISO	The International Organization of Standardization
ISP	Image signal processing pipeline
ITU	International Telecommunications Union
IVC	Images and video-communications database
JND	Just noticeable difference 0.75 proportion points on a psychometric function, where 75 % of the observers

	evaluate the stimulus to be greater than the comparison stimuli.
JPEG	Joint Photographic Experts Group
K	Kelvin
LCD	Liquid-crystal display
LIVE	Laboratory for Image & Video Engineering, University of Texas Austin
LIVE mobile	Laboratory for Image & Video Engineering Mobile Video Quality Database
LIVE(MDIG)	Laboratory for Image & Video Engineering, Multiple Distorted Image Database
LP/PH	Line pairs / picture height
LPC	Image sharpness assessment based on local phase coherence
lux	SI unit of illuminance used as a measure of the intensity, as perceived by the human eye
MDS	Multidimensional scaling, a statistical method for visualizing levels of similarity on abstract Cartesian space
MICT	Image database from Toyama University
MMSP (SVD)	Scalable Video Database, by Multimedia Signal Processing group
MOS	Mean opinion score
MTF	Modulation transfer function, a technical measure of sharpness and resolution of an imaging system
NIQE	Natural Image Quality Evaluator, a NR-IQA algorithm
NJQA	No-reference IQA for JPEG Images
NLP	Natural language processing
NR	No-reference
NR-IQA	No-reference Image Quality Assessment algorithm
NSS	Natural Screen Statistics is an application of the statistical regularities related to scenes
NYU Packet Loss Database	Packet Loss Video Database by New York University Video Lab
NYU Video Database	Video Database by New York University Video Lab
OECF	Opto-electrical conversion function
PC	Paired comparison
PCA	Principal component analysis
PSF	Point-spread function, describing the response of an imaging system to a point source or point object. The

	degree of spreading (blurring) of the point object is a measure for the quality of an imaging system.
px	Pixel
QBU	Question builder unit in the VQone toolbox
QCIF	Quarter Common Intermediate Format, referring to a video resolution of 176 x 144 pixels
QoE	Quality of experience
RR	Reduced reference
RR-IQA	Reduced reference image quality assessment algorithm
SAMVIQ	Subjective assessment method for video quality
SDSCE	Simultaneous Double Stimulus Continuous Evaluation
SFR	Spatial frequency response
SNR	Signal to noise ratio
SPSS	Statistical Package for the Social Sciences
SQS	Standard quality scale, the primary multivariate standard that can be used to derive an SRS yardstick in the Quality Ruler method
sRGB	Standard red-green-blue color space
SRS	Standard reference stimuli that observers use as a ruler to evaluate images in the Quality Ruler method
SSCQE	Single stimulus continuous quality evaluation
Sse	Sum of squared errors
SSIM	Structural similarity index metric
TID2008	Tampere Image Database 2008
TID2013	Tampere Image Database 2013
TUM	Technical University of Munich
VCX	Valued Camera eXperience.
VQA	Video quality assessment algorithm
VQEG	Video Quality Experts Group
VQEG FR-TV	Full Reference Television video database by Video Quality Experts Group
VQEG HDTV	High-Definition Television video database by Video Quality Experts Group

GLOSSARY

Avisynth	Tool for video post-production
Chi-squared distribution	Probability distribution used in statistical testing
Chroma	The colorfulness relative to the brightness of a similarly illuminated area that appears to be white
ETDRS chart	Early Treatment Diabetic Retinopathy Study Vision chart
Farnsworth D-15	Color vision and blindness arrangement test of 15 color plates
FinnWordNet	Lexical database for Finnish, a derivative of the Princeton WordNet
Gamma	A nonlinear operation used to encode and decode luminance values in imaging systems
Gretag Macbeth chart	Color calibration target consisting of a cardboard-framed arrangement of 24 squares of painted samples.
HuffYUV	Lossless video codec
Mahalanobis distance	Multi-dimensional generalization of the measure of how many standard deviations away a point is from the mean of the distribution
Matlab	Matrix Laboratory, a computing environment and programming language by MathWorks, Inc.
Photospace	Statistical method of describing the picture-taking frequency as a function of the subject illumination level and the subject-to-camera distance.
Qualinet	European Network on Quality of Experience in Multimedia and Services
Quality Ruler	A subjective image quality evaluation method where observers match the quality of the test items against a yardstick of ordered univariate reference images
Triplet comparison	Variation of the paired comparison method, where instead of two stimuli, the observers needs to compare three stimuli at a time
Venn diagram	A diagram that shows all possible logical relations between a finite collection of different sets
VirtualDub	Video capture and processing utility

1. INTRODUCTION

Do you know how many imaging devices you have at home? You probably have a smartphone with one, two, or even five or more cameras on it. Then, there is your laptop, tablet, television, gaming console, robotic vacuum, doorbell and security system. In 2000, Kodak estimated that consumers worldwide took approximately 80 billion photos in that year alone. Given that cameras have become a must-have standard feature in mobile phones, almost everyone takes photographs or records videos. A market research firm, InfoTrends, estimated that consumers had taken 1 trillion digital photos in 2015 and 1.2 trillion digital photos in 2017. The growth has been exponential, and it has been estimated that 10 percent of all photos ever taken since the invention of the camera in 1826 were taken during the last twelve months (Heyman, 2015). Images are everywhere, on billboards, art galleries, social media, news, television, a portrait of your loved ones decorating your desk or in a family album; they simply have become parts of our lives. Our memories and emotions are often preserved in imagery, and we cannot overestimate the importance of images as a means to transmit information and thoughts. With the advent of the internet and mobile phones, we communicate with images more than ever. Instagram, for example, had more than 1 billion monthly active user in 2018 (Instagram Corporation, 2019). Over 500 hours of videos are uploaded to YouTube every minute (Clement, 2019). This is all while the resolution and quality of the uploaded content has increased from 144p QCIF videos to 4K and even 8K videos (Kokaram, Foucu, & Hu, 2016). This explosion in visual content has created new demands to understand image quality and how people perceive images. Why do images convey information so well and why does the saying *'an image is worth a thousand words'* seem to be valid. Why do some images elicit emotions better than others despite depicting equally emotional content? What is the role of image quality in all of this?

The purpose of this dissertation is to evaluate the instruments available for image quality evaluation, develop new methods for subjective image quality evaluation and provide image and video databases for the assessment and development of image quality assessment (IQA) algorithms. As the topic of image quality has wide multidisciplinary relevance, this thesis has also combined different approaches by focusing on image quality as a psychological phenomenon and its relation to technical measurement.

1.1 IMAGE QUALITY AS A PSYCHOLOGICAL CONSTRUCT

One definition of image quality is related to image fidelity, particularly to perceptual fluency (Reber, Schwarz, & Winkielman, 2004). Images with clear perceptual fluency are preferred, as they can convey message better and are easier to interpret by the viewer. However, preference does not always follow fidelity (Fedorovskaya, de Ridder, & Blommaert, 1997), suggesting that there are also other processes involved. The dilemma of why something is preferred over another and why individual differences are so wide in preference, hence the saying ‘Beauty is in the eye of the beholder’, has plagued philosophers’ minds for centuries. It is therefore no surprise that the study of experimental aesthetics was also one of the earliest areas in psychology.

In 1876, Gustav Fechner published his *Vorschule der Aesthetik (Preschool of Aesthetics),* where he postulated that aesthetics as a science must proceed by employing empirical data to develop aesthetic theories. He hypothesized that the perception of aesthetic pleasure can be empirically comprehended as a result of the characteristics of the subject and the nature of the object (Fechner, 1876). Fechner not only raised the topic as a philosophical debate but also provided methods and theory for the measurements of the relation between sensation and perception in the form of psychophysics (Gescheider, 1985). It can be argued that subjective image quality assessment methods are also strongly rooted in psychophysics (Engeldrum, 2000; Keelan, 2002; Winkler, 2005). Although the scientific study of image quality shares much of its origin with experimental aesthetics, it is also a subsection of the highly multidisciplinary science of quality of experience (QoE), consisting of the primary disciplines of vision science (To, Lovell, Troscianko, & Tolhurst, 2008), color science (Yendrikhovskij, de Ridder, Fedorovskaya, & Blommaert, 1997) as well as the computational sciences (Dodge & Karam, 2019; Redi, Zhu, de Ridder, & Heynderickx, 2015) and behavioral sciences (Augustin, Wagemans, & Carbon, 2012; Leder, Belke, Oeberst, & Augustin, 2004; Leisti, Radun, Virtanen, Halonen, & Nyman, 2009; Nyman, Radun, Leisti, & Vuori, 2005; Tinio, Leder, & Strasser, 2011). It is clear that image quality has great relevance to various disciplines, and its importance to industry is undeniable. It is therefore no surprise that many definitions of image quality have been devised.

QoE and image quality are defined differently in various sources. The most ambitious effort to create a comprehensive definition of QoE has likely been given by Qualinet, the European network on Quality of Experience for multimedia systems and services. The working definition of QoE was created by 49 researchers representing 18 European countries: *“Quality of Experience (QoE) is the degree of delight or annoyance of the user of an application or service. It results from the fulfilment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user’s personality and current state“* (Le Callet, Möller, & Perkins, 2012). The researchers themselves note that the current definition does not address the degree of success achieved by the artist to convey the intended message but rather what influence does the technical system or processing have on the artist’s work. The International Imaging Industry Association (I3A) Camera Phone Image Quality (CPIQ) Initiative group defined image quality to be *the perceptually weighted combination of all visually significant attributes of an* *image when considered in its marketplace or application* (I3A, 2007), whereas Janssen and Blommaert defined the quality of an image to be "*the degree to which the image is both useful and natural"* (Janssen & Blommaert, 1997). Engeldrum, on the other hand, described image quality to be "*the integrated perception of the overall degree of excellence of an image"* (Engeldrum, 2004b) and Keelan characterized image quality as "*the impression of its merit or excellence, as perceived by an observer neither associated with the act of photography, nor closely involved with the subject matter depicted"* (Keelan, 2002). This variation in definitions reflects various time periods and application areas but also shows how the context and research area affect the definition.

Image quality development and research can be approached from two different perspectives: from bottom-up and top-down perspectives (Yendrikhovskij, MacDonald, Bech, & Jensen, 1999). The former starts with objective measurements of the device parameters, i.e., the signal-to-noise ratio (SNR), following with estimations of the magnitude of psychophysical sensations that they introduce, i.e., graininess. In this view, the absence of visible distortions creates high quality – image fidelity is image quality. Equating fidelity with quality has been challenged, for example, by Nyman et al. (2005), and some studies even show observers actually preferring certain distortions such as oversaturated colors (Fedorovskaya et al., 1997). To explain these phenomena, a top-down view is presented (Janssen & Blommaert, 1997). It in contrast argues the view that images are processed as information about the outside world, not as signals. Accordingly, it is argued that image quality can be defined as the degree to which the image can be successfully exploited by the observer. A concept that does not contradict either view suggests that perceptual and conceptual processing fluency could be an intrinsically pleasurable experience (Reber, Schwarz, & Winkielman, 2004). Images with clear perceptual fluency are preferred because they can convey a message better and are easier to interpret by the viewer. However, this processing fluency cannot explain why abstract art is perceived as aesthetically pleasing. This led to a dual-processing perspective to the processing fluency theory, where abstract art, with its low processing fluency, would introduce aesthetic pleasure through cognitive enrichment, while natural scene images would be processed mostly at an automatic level, in which clear processing fluency would be preferred (Graf & Landwehr, 2015). It has also been suggested that people have an understanding of images simply being representations of the scene that they depict and are preferred by their degree of artistic value, where image quality is one significant factor (Tinio et al., 2011). It can also be claimed that, being exposed to thousands of images during their life, people become accustomed to assessing image quality and at the least have certain expectations on what they consider good image quality. However, as imaging and display devices further develop and new technologies emerge, expectations will change as well.

1.1.1 IMAGE QUALITY ATTRIBUTES

Image quality can also be conceptualized as a combination of preferential attributes, also known as *nesses such as sharpness or colorfulness. The International Organization of Standardization (ISO) defines the preferential attribute as an attribute of image quality that is invariably evident in an image and for which the preferred degree is a matter of opinion, depending upon both the observer and the image content (ISO, 2005a). These preferential attributes are weighted and summed to create an overall model of image quality (Bech et al., 1996; Engeldrum, 1999, 2004b; I3A, 2007; Janssen & Blommaert, 1997; Keelan, 2002; Yendrikhovskij, Blommaert, & de Ridder, 1999). This definition has the benefit of combining the views from multidisciplinary stakeholders approaching image quality from different directions. The summation and weighting of the preferential attributes or elements can be viewed as reflecting the cognitive-affective process of the viewer. For example, Berlyne (1972) suggested that preference is formed from the combination of pleasingness, interestingness, liking and complexity. O'Hare and Gordon (1977) linked realistic-unrealistic, clear-indefinite and symmetrical-asymmetrical dimensions to the preference of paintings. The concept of the summation of image elements and scene statistics is also useful from a technological point of view when developing image quality assessment algorithms such as the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE), which is a natural scene statistic-based distortion-generic blind/no-reference (NR) quality assessment algorithm (Mittal, Moorthy, & Bovik, 2012), and Video BLIINDS, a natural scene statistic model-based approach to the no-reference/blind video quality assessment problem (Saad, Bovik, & Charrier, 2010). Both technical and psychological approaches often utilize some type of summation of image elements or preferential attributes to create a model for image quality perception. The common understanding is that preference can be broken up into smaller elements, which can then be measured, summed and weighted.

Perhaps because of its multidisciplinary relevance, the terminology of QoE and image quality has been poorly defined (Augustin et al., 2012; Virtanen, Nuutinen, & Häkkinen, 2019). Researchers lack consensus on the most fundamental attributes of image quality and audio-visual quality. For example, usefulness and naturalness were considered defining attributes by Janssen (2001). Sharpness was considered one of the most critical attributes utilized in image quality models (Engeldrum, 1999). Yendrikhovskij et al. (1999) considered naturalness, visibility of details, brightness rendering and chromatic rendering as critical to color television displays. A more general foundation for the terminology of aesthetic word use was given by Augustin et al. (2012), who examined the aesthetic word use with eight different object classes, therein showing an interplay of generality and specificity in aesthetic word usage.

A study by Nyman et al. (2010) demonstrated how images with low quality ratings were characterized by different terminology compared to images with high quality ratings, suggesting a 'subjective paradigm shift' in the subjective decision-making space as a function of preference. In other words, images of low quality are evaluated with a different set of rules and terms than images of high quality. The same concept also applies to printed images, as demonstrated by Leisti et al. (2009), who further classified the terminology on image quality to have two levels: low level and high level. The most important low-level attributes were the brightness of color, sharpness, graininess, brightness, color quality, gloss, contrast, and lightness. High-level attributes, on the other hand, were used to funnel the importance of the low-level attributes and consist of realism, naturalness, clarity, depth, and aesthetic associations.

1.1.2 METHODS OF SUBJECTIVE IMAGE QUALITY EVALUATION

It is a recommended practice to follow standards that define and specify detailed viewing conditions and calibrations for the presentation of the stimuli (ISO, 2005a, 2009; ITU, 2008a, 2012b, 2016; Streijl, Winkler, & Hands, 2016). Standards enable direct comparison of the results between different research groups and laboratories, which facilitates fruitful discussion within the research community. The viewing conditions and environment should be controlled, and the tests should be conducted in a room devoted to that purpose. The environment should be a non-distracting, comfortable and quiet, and people not involved in the experiment should not be present (ISO, 2009; ITU, 2016). For example, walls, ceilings, floors, and other surfaces in the space where the assessments are conducted should be a neutral matte gray with a reflectance of 60 % or less (ISO, 2009). However, when viewing images or videos on a display, the requirements, especially for ambient illumination levels, change (ISO, 2005a; ITU, 2008b, 2012b). For example, the environment could simulate a common living room for television viewing, with the respective screen size and viewing distance set to match that use case (ITU, 1997, 1998a, 1998b, 2012b, 2012a). Display requirements and calibration are also important factors to consider. A recommended practice is to follow the parameters given in standards so that the results can be replicated and compared by others (ISO, 2005a, 2009; ITU, 2008a, 2012b).

ITU-T P.913 (ITU, 2016) also discusses sampling techniques for the test subjects. The most common method in image quality studies is convenience sampling, which simply means that sampling is based on the availability and convenience of the researcher, for example, university students, until a sufficient number of subjects has been acquired. Probability sampling, on the other hand, dictates that all the relevant elements from a population should be included in the selected sample. Probability sampling can be achieved with various techniques depending on the goals of the study. For example, stratified random sampling divides the population into smaller groups based on a set of characteristics deemed relevant for the study such that subjects from each group are represented in the final sample.

It is crucial to understand the population from which the desired results are to be drawn. ITU-T P.913 (ITU, 2016) suggests to consider at least the following when selecting the sample for an experiment: 1. Use case specificity: does the case have a very specific implementation such as video streaming QoE? 2. Population segment specificity: Do we need to have expert observers who know what to look for and where to look for regarding distortions or does the sample need to represent average consumers? 3. Geographical location: Is the population drawn to a single location or does it require multiple locations? These considerations might seem quite straightforward at first but should be thoroughly contemplated. For example, using only expert observers can also distort the results, as they might weight the effect of distortions to image quality differently than naïve observers would. After the criteria above are considered, ITU-T P.913 also recommends aiming at a 50:50 gender distribution and balanced age distribution unless otherwise required by the experimental design (ITU, 2016). A balanced age distribution can depend on the focus of the study. Do we need to represent the whole population or are we simply interested in 20- to 30-year-olds? It is also recommended that observers be checked for normal vision characteristics insofar as they affect their ability to carry out the assessment task. This means confirmation of normal color vision and normal or corrected-to-normal visual acuity applicable to the viewing distance in the experiment. With audio-visual stimuli, normal hearing aptitude should also be screened for. (ISO, 2005a; ITU, 2007, 2016). Finally, the criteria for selecting observers and notable characteristics of the observer group as a whole should be reported with the results (ISO, 2005a; ITU, 2007, 2016).

Industry standards are periodically reviewed and updated because many of the published standards represent use cases that can be outdated. For example, ITU BT. 500-13 concerns studio quality videos, viewed on CRT screens in a living room environment (ITU, 2012b). However, the subjective methodology suggested by these standards is often still valid and useful for various needs. These standards have the benefit of being well documented, widely accepted, thoroughly tested and extensively replicated. Industry-mandated standards are a good reference on how to build up the lab environment and what things should be considered. It is equally important that academic researchers do not feel bound by industry-mandated standards of conduct regarding any type of study (Moorthy, Choi, Bovik, & de Veciana, 2012). The standards and recommendations create a good baseline from where to start; however, they should not be considered as restrictive for future method development or research questions.

With a few exceptions and variations, the methods for subjective assessments presented in various standards and recommendations can be divided into the following categories. The *media category* classifies methods based on whether they are used for image quality or video quality assessments. Some methods can also be applied for both. The *task category* can be divided into two groups. A rating task utilizes some type of scale that the observers use to evaluate the stimuli. In the comparison task, the observers select the stimuli that they prefer among multiple stimuli. The *reference category* dictates whether the method includes some type of reference stimuli to anchor the ratings. In addition, it is possible to use no reference at all or to explicitly present a reference for the observers. There is a third group, hidden reference, where the reference stimuli are presented to the observer as one of the assessed stimuli. The *stimuli category* divides the methods by how many stimuli are shown simultaneously to the observer. This can also include temporal presentation, for example, presenting the stimuli in pairs one after the other. The *evaluation category* divides the methods into absolute assessment, degradation assessment and continuous assessment. For absolute assessment, the observer evaluates the stimuli for its perceived quality. In the degradation assessment, the observer evaluates the amount of degradation the stimuli presents versus a reference. The continuous assessment gives a time stamped continuous rating by having the observer move a slider according to the temporal fluctuations of the quality of the stimuli. See table 1.

Table 1. *Table 1: Overview of recommended methods described in various standards.*

Method	Acronym	Standard	Media	Task	Stimuli	Evaluation	Reference
Subjective Assessment Method for Video Quality evaluation	SAMVIQ	ITU-R BT.1788	Video	Rating	One	Absolute	Explicit & Hidden
Double-Stimulus Continuous Quality Scale	DSCQS	ITU-R BT.500-13	Both	Rating	Two	Absolute	Hidden
Single-Stimulus Continuous Quality Evaluation	SSCQE	ITU-R BT.500-13	Video	Rating	One	Continuous	No
Double-Stimulus Impairment Scale[1]	DSIS	ITU-R BT.500-13	Video	Comparison	Two	Degradation	Explicit
Degradation Category Rating[1]	DCR	ITU-T P.910	Video	Comparison	Two	Degradation	Explicit
Simultaneous Double-Stimulus Continuous Evaluation	SDSCE	ITU-T P.910	Video	Rating	Two	Continuous Degradation	Explicit
Absolute Category Rating with Hidden Reference	ACR-HR	ITU-T P.910	Both	Rating	One	Absolute	Hidden
Absolute Category Rating	ACR	ITU-T P.910	Both	Rating	One	Absolute	No
Paired Comparison	PC	ITU-T P.910	Both	Comparison	Two	Absolute	No
Triplet Comparison	-	ISO 20462-3	Image	Comparison	Three	Absolute	No
Quality Ruler[2]	QR	ISO 20462-2	Image	Comparison	One	Absolute	Ruler

1. Identical methods

2. In QR, the observer matches the stimuli against a set of gradually degraded ruler images

The standards list other possible rating tasks such as similarity and performance-based methods (ITU, 1990, 2012b). For example, in the latter, the accuracy and speed of an externally directed performance task (reading, searching, etc.) are used as a measure (ITU, 2012b). These measures are not, however, related to image quality evaluation and are therefore omitted from this list.

Each method of Table 1 is presented and evaluated in greater detail in the following sections. First, the simplest form of the rating methods, the absolute category rating (ACR), is presented, from which other rating methods have been derived. The second method presented is paired comparison (PC), which is the simplest form of comparison methods. These two methods will provide a general understanding of the differences between the two approaches with their advantages and disadvantages before the variations from each method are listed.

1.1.3 ABSOLUTE CATEGORY RATING (ACR)

The ACR is probably the easiest method to implement, as described in ITU-T Rec. P.910 (ITU, 2008a). The recommendation considers the use of the method only from the viewpoint of video quality assessment; however, it can also be used with images as well. The ACR method is a single-stimulus method, where the observers' rate images or video clips one at a time, and no reference is presented for the observers. Observers use a 1-5 rating scale, with discrete categorical labels for quality: *Excellent*=5, *Good*=4, *Fair*=3, *Poor*=2, *Bad*=1. The evaluations are then averaged to create a mean opinion score (MOS) for each evaluated system such as a video codec or an image capture device. A sufficient number of replications can also be obtained by repeating the same test stimulus at different occasions during the test. The benefit of the ACR method is that it is easy to set up and provides instructions to the observers, and its results are simple to analyze and communicate. The ACR method can be easily modified to evaluate specific quality dimensions, such as brightness and sharpness, instead of overall quality.

If higher discriminative power is required, a nine-level scale may be used, where the categorical labels for quality are *Excellent*=9, *Good*=7, *Fair*=5, *Poor*=3, *Bad*=1. Annex B in the ITU P.910 standard also further considers an 11-level scale and a graphical continuous 0-100 scale, where categorical labels are only shown at the endpoints, which should reduce the bias due to the interpretation of the category labels by the observers (ITU, 2008a). The continuous scale is claimed to be superior to the 5-category judgment scale because it allows observers to indicate finer gradations in visual quality (Seshadrinathan, Soundararajan, Bovik, & Cormack, 2010). There is also a tendency for the observer to use each of the categories (except sometimes the two end categories, which may be held in reserve), regardless of the adjectival descriptors associated with them (ISO, 2005a).

As in any method that uses scales, respondents vary in their usage of the scale. It is irrelevant whether the scale is continuous or categorical. Common patterns include using only the middle of the scale or using the upper or lower end, which can impart biases to many of the standard analyses conducted with rating data, including regression and clustering methods, as well as the identification of individuals with extreme views. A standard procedure for addressing scale usage heterogeneity is to transform the data into a z-score that centers each respondent's data by subtracting the overall mean over all questions and dividing by the overall standard deviation (Sheikh, Sabir, & Bovik, 2006; van Dijk, Martens, & Watson, 1995); however, other methods have also been suggested (Rossi, Gilula, & Allenby, 2001). Because the ACR method does not have any reference (hidden or explicit), its results can be difficult to compare across different laboratories and institutions. Without a reference to use as an anchor for aligning the scale across different locations, differences in observer population, scale use, test material, etc. make it difficult to compare the results. In addition, as people's expectations change over time when technology improves, the results can become incomparable with new studies after a certain period of time.

It is also not straightforward to translate the names of the scale categories into different languages. In doing so, the inter-category relationship can become different from that in the original language (ITU, 1990, 2008a). Some studies have shown that the mental distance between the *Excellent* and *Good* categories is not equal to the mental distance between the *Poor* and *Bad* categories (Teunissen, 1996). People perceive and use these labels differently, which can affect their evaluations. It is possible to counter both of these concerns by giving adjective labels only to the endpoints of the scale. When there are no adjectival categories between the endpoints, the varying mental distances between categories do not affect the results as much, and it would be perceived as more continuous, even if the numerical length of the scale remains the same.

The MOS has become the "de-facto" metric of perceived quality. The benefit from this has been the raised awareness of the importance of the perceptual aspect of quality. It is helpful to have a clear and easy-to-understand quality indicator that has widespread acceptance. Unfortunately, there has not been much consideration on the limitations and restrictions of the subjective experimental design. MOS is often reported without sufficient understanding of how the data have been obtained and without attention to the selected method's accuracy, reliability, or applicability (Streijl et al., 2016).

Condensing the results of subjective assessment into MOS values can hide valuable information related to inter-user variation. Providing a standard deviation with the MOS values does not remedy this problem completely because two very different assessment distributions can "hide" behind the MOS (Hoßfeld, Heegaard, Varela, & Möller, 2016). The standard deviation is typically highest around the middle of the MOS range and decreases toward the ends of the scale. This behavior can be observed for most experiments, independent of the specific rating scale used (Virtanen, Nuutinen, Vaahteranoksa, Oittinen, & Häkkinen, 2015; Winkler, 2009; Winkler & Dufaux, 2003). In theory, scales with higher granularity should reduce the standard deviations of the MOS; in practice, however, these differences turn out to be insignificant (Winkler, 2009). In addition to the standard deviation, providing the skewness and kurtosis measures can give a better picture of the distribution behind the MOS value.

1.1.4 PAIRED COMPARISON (PC)

The paired comparison method is presented in ITU-T P.910, ITU-R BT.1082-1, ISO 20462-1 Annex B standards (ISO, 2005a; ITU, 1990, 2008a). It is effectively a method whereby the test stimuli are presented in pairs, and the observer is required to make a forced judgment between the two stimuli using a specific criterion, e.g., preference or sharpness, under study. This method of forced judgment between two stimuli is also known in the psychophysical literature as the two-alternative forced choice (2-AFC) method. In the case of videos, the presentation is often temporally separated, showing one video after the other in random order, and spatially separated pairs can also be used with images. PC was one of the earliest methods used in experimental psychology and in the study of aesthetics. In his book, *Vorschule der Aesthetik*, Fechner suggested that the pleasantness of two objects could be studied by having observers choose the object that is more pleasant (Fechner, 1876). Later, Thurstone published a study on the law of comparative judgment applied to paired comparison, where the method facilitated more thorough theoretical analysis for the data it could provide (Thurstone, 1927).

The items under tests (A, B, C, etc.) are generally combined in all possible combinations: AB, BA, CA, etc. Thus, all pairs in a sequence should be displayed in both possible orders (e.g., AB and BA). When the presentation order is considered, the number of sample combinations for paired comparison N is expressed by

$$N = n(n - 1) \tag{1}$$

Where n is the number of samples and n = 2, 3, 4, 5, etc. However, especially with spatial presentations whereby images are shown side by side, the order can often be ignored if the position, left or right, is randomized. In these cases, the number of sample combinations for paired comparison N is expressed by

$$N = n(n - 1)/2 \tag{2}$$

As the number of pairs increases exponentially, the method is best suited for situations where the number of tested items is small. A binary sorting tree method for selecting which pairs to compare based on previous comparisons has been suggested as a way to increase the time efficiency of the method (Farrell, 2001). Another option is to reduce the number of test stimuli by using faster rating methods, such as ACR or DCR, first and then using PC on those items that have received approximately the same rating (ITU, 2008a). The comparison data can also be transformed into an interval scale with a technique based on Thurstone's Law of Categorical Judgment (Torgerson, 1958). A method for this conversion is given in ISO 20462-2 Annex E (ISO, 2005b).

ISO 20462-2 Annex F provides a method for converting the data into a just noticeable difference (JND) measure between two or more stimuli. In psychometrics, the JND is defined as the 0.75 proportion points on a psychometric function, where 75 % of the observers evaluate the stimulus to be greater than the comparison stimuli (Gescheider, 1985). When the probability of choosing between two stimuli is 50 %, they can be considered equal, as either one would be chosen by chance.

The PC method has a high discriminatory power, which is of particular value when several of the test items are of equal quality (ITU, 2008b). It is therefore an especially good method for situations whereby the perceived difference is small or there is a need to determine if the difference is strong enough to elect a noticeable difference. It was also found to be the most accurate method in a study comparing four subjective methods for image quality assessment (Mantiuk, Tomaszewska, & Mantiuk, 2012). However, if the stimulus difference exceeds approximately 1.5 JNDs, the magnitude of the difference cannot be directly estimated reliably because the response saturates as the proportions approach unanimity, e.g., one stimulus is selected 100 % of the time over the other in paired viewings (ISO, 2005a). The exponential growth of comparisons as a function of the number of test stimuli reduces the utility of the 2-AFC PC method. Despite these drawbacks, it is a powerful method when used with suitable test material and research questions.

1.1.5 TRIPLET COMPARISON

The triplet comparison method was introduced in ISO 20462-2 Triplet comparison method (ISO, 2005b). The method is presented as a two-step process, where the first step is to rank the images depicting the same scene into three categories: *"favorable"*, *"acceptable"*, or *"unacceptable"*. In the second step, the observers see three images depicting the same scene and are instructed to rate them in order of preference. The method is a forced choice rating method, where the option for giving the same rank for two or more images is prevented.

The randomization protocol in the triplet comparison method uses balanced incomplete block (BIB) design, where each stimulus is paired against each other at least once for all of the triplet combinations (Burton & Nerlove, 1976; ISO, 2005b). For example, with items 1 to 9, combinations without duplication can be achieved with just 12 triads: (1, 2, 4), (4, 5, 7), (7, 8, 1), (2, 3, 5), (5, 6, 8), (8, 9, 2), (1, 3, 6), (4, 6, 9), (7, 9, 3), (1, 5, 9), (4, 8, 3) and (7, 2, 6). Without balancing the blocks and preventing duplicate pairs, nine items would create a complete set of 84 triads, which would create an exhausting experiment for the observers.

Using triplet comparison instead of PC has the benefit of reducing the experiment time, as it reduces the number of sample combinations. The number of sample combinations for triplet comparison N is expressed by

$$N = n(n-1)/6 \tag{3}$$

Where n is the number of samples and n = 2, 3, 4, 5, etc. When comparing the number of sample combinations from PC with Equation 2 against the sample combinations from the triplet comparison with Equation 3, the triplet comparison with BIB design reduces the number of sample combinations to one third of that of PC, as the divisor is six in Equation 3 rather than two, as in Equation 2. In the previous 9 sample examples, the sample combinations can be presented with just 12 triplets, whereas it would require 36 pairs when using the PC method. However, not all sample sizes are valid for balanced design triplet comparison without duplicated pairs, and the number of samples is restricted to n= 7, 9, 13, 15, 19, 21, and 27. Sample sizes greater than 27 are possible; however, 27 samples already create 117 triads.

The preceding task of ranking into three categories, *"favorable"*, *"acceptable"*, and *"unacceptable",* is used to reduce the quality variation and number of stimuli for the following triplet comparisons in the next step. As with PC, the triplet comparison method works well in situations whereby there are few samples with quite small quality variations among them. If there are only a few items in the experiment, the sorting task can be omitted. In the first step, all stimuli are simultaneously ranked by the observer, which may be impractical with softcopy or projected image display, and can place stringent requirements on the size of an observation area, which should provide uniform and equivalent viewing conditions ISO 20462-1(ISO, 2005a).

As with PC, the method works well with few samples within a similar quality range; however, the sample combinations still grow exponentially, although to a lesser degree, with triplet comparison. The two methods were compared in a study presented in Annex A of ISO 20462-2, where they were found to be similar with respect to their consistency and accuracy. However, the desirable nature of the triplet comparison decreases the level of stress on the observer due to reduced assessment time, which suggests that the triplet comparison method has the potential to achieve consistent and accurate results. Triplet comparison data can also be transformed into an interval scale with a technique based on Thurstone's Law of Categorical Judgment (Torgerson, 1958). A method of converting the results into the JND scale is given in (ISO, 2005b).

1.1.6 ABSOLUTE CATEGORY RATING WITH HIDDEN REFERENCE (ACR-HR)

A modification to the ACR method was presented in ITU-T Rec. P.910 (ITU, 2008a). In the ACR-HR method, the reference image or video is "hidden" among the test stimuli, and observers evaluate it just like any other test item. The rating task for the observer remains the same as in the ACR method; however, during analysis, a differential quality score (DMOS) can be computed for each test stimulus by comparing it to the corresponding (hidden) reference. As with the ACR method, a sufficient number of replications can be obtained by repeating the same test stimulus at different occasions during the test.

There is also the same possibility to adjust the rating scale or use a graphical continuous rating scale as in the ACR method. The method can be easily adjusted to evaluate specific quality dimensions. Such dimensions may be useful for obtaining more information on different perceptual quality factors when the overall quality rating is nearly equal for certain systems under test but when the systems are clearly perceived as different.

ACR-HR has the advantages of ACR with respect to presentation and speed, and the use of a hidden reference can remove some biases due to the scene or the observers liking or disliking certain content. However, the method suffers from the same disadvantages of categorical scaling as the ACR method described above. These can be mitigated to a certain extent by using a continuous scale.

1.1.7 DEGRADATION CATEGORY RATING (DCR) AND DOUBLE STIMULUS IMPAIRMENT SCALE (DSIS)

The degradation category rating (DCR) presented in ITU-T Rec. P.910 (ITU, 2008a) is also known as the double-stimulus impairment scale (DSIS) method described in ITU-R BT.500-13 (ITU, 2012b). The DCR includes paired viewing of the video clips, where each clip is preceded by a corresponding reference clip. Observers rate the level of impairment compared to the reference using the 1-5 rating scale with discrete categorical labels for impairment: *Imperceptible*=5, *Perceptible but not annoying*=4, *Slightly annoying*=3, *Annoying*=2, and *Very annoying*=1. A nine-level scale version for the DCR method is given in ITU-T Rec. P.910 Appendix V. A variation of the DCR method is to display the reference and the test sequence simultaneously on the same monitor so that the reference is located on either side of the stimuli.

ITU-T P.910 (ITU, 2008a) recommends DCR be applied in high-quality system evaluation. The discrimination between imperceptible/perceptible but not annoying categories might bring some added value when compared against the original reference sources. However, the DCR method still suffers from the same issues as other categorical ratings such as the ACR method presented above. The mental distance between the adjectival categories *annoying* and *slightly annoying* could differ between observers, as they may interpret the terms differently. Translating the categories will also introduce another layer of variation to the results that can make comparison between laboratories more difficult.

1.1.8 DOUBLE STIMULUS CONTINUOUS QUALITY SCALE (DSCQS)

Presented in ITU-R BT.500-13, the DSCQS is a method whereby observers are presented with a series of image or video pairs in a random order (ITU, 2012b). Each pair is also presented in internally random order and consists of two versions of the same stimulus, where one version is the original source stimulus without any impairment and the other version contains some process or impairment manipulation under study. The observers rate the perceived quality for both stimuli using a continuous 0-100 quality scale and go through all the required combinations of samples. There are two variants of the DSCQS method. The first variation is conducted by a single observer at a time. For each presentation of a stimulus pair, the observer is free to view both stimuli until he can assign a mental measure of quality associated with each stimuli for rating. Variant two uses more than one observer at the same time, and the stimulus pairs are shown one or more times while the results are recorded. The number of repetitions of the pairs is dependent on the duration of the stimuli. For still images, a 3-4 second viewing with five repetitions (voting during the last two repetitions) may be appropriate, while a 10 second video sequence with two repetitions (voting during the last viewing) may be appropriate.

The DSCQS is an interesting combination of the forced choice PC and the continuous rating task. The use of a continuous scale instead of categorical adjectival labels reduces the risk of observers using the scale differently because of variations in their interpretations of the category terms. With the added granularity of a scale, DSCQS can be used with stimuli having wider quality variation compared to the forced choice PC. With PC, after the stimulus difference becomes too large, the proportions approach unanimity, e.g., one stimulus is selected 100 % of the time over the other stimulus in paired viewings. This method still suffers from the same problems with variations in scale usage as any other method. Because the observers need to have two adjacent rating scales for rating both stimuli at the same time, the task can become more demanding than giving a single rating. Some studies have found that there is a risk of misplaced ratings when the observers confuse which rating is associated with which stimulus in the pair (Pinson & Wolf, 2003). These obvious outlier ratings can fortunately easily be screened out by examining the data; however, any missing data is still unfortunate.

1.1.9 SAMVIQ SUBJECTIVE ASSESSMENT METHOD FOR VIDEO QUALITY

Presented in ITU-R BT.1788, the SAMVIQ quality evaluation method is derived from the DSCQS method (ITU, 2007). In this method, the viewer is given access to several processed versions of a video sequence. They randomly select which version they want to view and perform their evaluation using a graphical user interface (GUI) and can go back, review and modify their ratings of each processed sequence as desired. They are also given access to an explicit, unprocessed reference that they can view at any time. The SAMVIQ method includes a hidden reference identical to the explicit reference. Each version of a sequence is displayed alone and rated using a continuous scale graded from 0 to 100, annotated by 5 quality categories (*Excellent*, *Good*, *Fair*, *Poor*, and *Bad*) spaced evenly on the scale.

The idea of including both explicit and hidden reference in the method is interesting. The ITU-R BT.1788 recommendation also notes that the explicit name "reference" can have an impact on some observers who then give the explicit reference the highest possible score, while the corresponding hidden reference is scored as something completely different even though the two sequences are identical (ITU, 2007). Having both explicit and hidden references could aid in screening the data for inconsistent observers and outliers. The decision to use adjectival labels along with the continuous scale is problematic because it introduces difficulties in translating the labels into different languages. Additionally, the mental distance between Excellent and Good might not be perceived as equal to the mental distance between Poor and Bad, as noticed by Teunissen (1996).

1.1.10 SINGLE STIMULUS CONTINUOUS QUALITY EVALUATION (SSCQE)

Presented in ITU-R BT.500-13, the SSCQE is designed for the evaluation of distortions that are scene dependent and time varying such as transmission distortions (ITU, 2012b). Even within short extracts of digitally coded video, the quality can fluctuate quite widely depending on the scene content, and impairments may be very short lived. Different video sequences can contain different amounts of spatial information (SI) and temporal information (TI) (ITU, 2008a). This variation between sequences can considerably affect the visibility and amount of impairment. This is true for compression schemes and concerns the error resilience behavior of digital transmission systems (ITU, 2012b). Using only a single rating at the end of a video clip will not capture this temporal variation in quality. The previous methods are also limited to short presentation durations for each video clip.

In SSCQE, the observer's task is to evaluate the perceived quality of a video by moving a graphical or physical continuous slider accordingly. The position of the slider is time coded and recorded. The participant's mean quality rating q can then be mapped as a function of time t, $q(t)$, where the quality can change over time depending on the scene and its time-varying distortions. Hence, the quality rating can be calculated for just a sequence segment, a quality parameter or for the entire test session depending on the needs of the study. However, the varying delay in different observer response times may influence the assessment results if only the average over a segment of the video is calculated.

One of the benefits of using a continuous quality evaluation is that it is not affected by human memory bias; for example, distortions at the end of the video can be more influential than distortions at the beginning of the video, and the task is to give a single overall rating to the whole video clip. As with any other method using a scale, the SSCQE is also vulnerable to individual differences in scale usage by the observers. Considering the temporal aspect of the task, observers can be inclined to postpone the use of endpoints of the scale in case something worse or better may later appear.

1.1.11 SIMULTANEOUS DOUBLE-STIMULUS CONTINUOUS EVALUATION (SDSCE)

In Appendix III of the ITU-T P.910, SDSCE is presented for evaluating effects for sparse impairments such as transmission errors on the fidelity of visual information (ITU, 2008a). This method derives from the single-stimulus continuous quality evaluation (SSCQE) from the ITU-BT.500-13 standard (ITU, 2012b). Here, the observers view two videos side by side, where one video is the reference and the other video is the test stimulus. The task is to evaluate the differences between the two videos by moving a graphical or physical continuous slider. When a test stimulus does not have any visible difference compared to the reference, the fidelity is considered 100 %, and the slider should be kept at that position. When visible degradations occur, the observer should move the slider accordingly to match the perceived fidelity on a scale from 100 to 0. The position of the slider is coded and recorded. The data can be plotted in a similar manner as in the SSCQE method, and the observer's mean fidelity rating can be calculated for a sequence segment, a quality parameter or for the entire test session depending on the aims of the study. However, the same varying delay in observer response times may influence the assessment results if only the average over a segment of the video is calculated.

1.1.12 QUALITY RULER

The quality ruler method was presented in the ISO 20462-3 standard, and it describes a psychophysical method that involves quality assessment of a test stimulus against a yardstick of ordered univariate standard reference stimuli (SRS) (ISO, 2005c). The standard reference stimuli (SRS) differ by increments of known numbers of JNDs. In the quality ruler method, observers select an image from a given set of SRS images that would correspond to the test image in terms of quality. In other words, they try to match the ruler SRS image to the test image so that they are of equal image quality. The ISO 20462-3 standard also makes an interesting claim that scene content and other properties of test images need not match those of the ruler in the quality ruler evaluation task. For example, one could use a ruler image depicting a portrait and try to match its quality with a landscape test image. Furthermore, the SRS would not need to have the same distortion as the tested images; thus, a ruler with varying levels of blur could be used to match its quality against test stimuli that vary in color distortions.

A primary multivariate standard, the standard quality scale (SQS), is the basis of the quality ruler method. The SQS was formed by showing subjects images varying in all aspects of image quality and asking them to rate the pictures using a magnitude estimation scale. Images with very similar rated scores are identified and grouped. These groups were then individually rank ordered for overall quality by both trained observers and representative consumers. This rank order was considered to represent multiple paired comparisons, and the probability of choosing one stimulus over the other was deduced from the rankings given to each stimulus by each observer (ISO, 2005c).

The outcomes were the 30 quality JND steps found in the SQS, where a value of zero corresponds to an image in which the principal subject is difficult to identify, and a grade of 30 falls within the range classified as excellent by consumers. In the original work concerning SQS development, the authors stated that this method combining magnitude estimation and rank ordering is comparable to PC (Keelan & Urabe, 2003).

The quality ruler requires the presence of a known set of SRS that acts as a ruler and that is calibrated against the SQS. Instructions for creating a sharpness-based SRS are provided in the ISO 20462-3 standard in Annex D. A set of images was produced by progressively blurring high-fidelity scene captures and rated by trained experts using the primary multivariate standard: SQS. The resulting blur steps were then linked to the system modulation transfer function (MTF), a technical measure of sharpness and resolution, with values that act as a link function enabling the creation of rulers with approximately known JND intervals (ISO, 2005c). An SRS set of images is also provided with a softcopy version of the quality ruler (Jin & Keelan, 2009).

The idea to use a set of images as a ruler in the rating alleviates many of the problems of using rating scales. The quality ruler method is innovative and provides a common yardstick anchored to a physical measure for all test laboratories across different locations and institutions. However, the MTF-based link function gives only an approximate JND step for the creation of a ruler image set. Different scenes act differently on the variation of the MTF. For example, a landscape with high-frequency information is likely to have stronger-than-average quality dependencies on MTF than some portraits.

1.2 IMAGE QUALITY FROM THE TECHNICAL PERSPECTIVE

1.2.1 TECHNICAL MEASURES WITH TEST CHARTS

Image quality is often approached strictly from the technical fidelity perspective. Several standards and recommendations have been created to characterize and measure various aspects of imaging devices using test targets, such as sharpness (ISO, 2017a), noise (ISO, 2017b), optical distortions (ISO, 2016), exposure (ISO, 2019), optoelectrical conversion function (OECF) (ISO, 2015), and color (ISO, 2017c), for characterizing imaging devices and measuring technical features that influence the perceived quality of the images that they produce. In addition to test target and measure-specific standards, recommendations are also made for specific use cases such as mobile phone image quality (IEEE, 2017) and automotive image quality (IEEE P2020 Working Group, 2018). The purpose of these objective test target measures is not to directly predict or model how observers would perceive and evaluate natural images – or the test targets themselves for that matter. These measures are used to characterize how an imaging system reproduces, distorts and manipulates signals captured under controlled conditions. Some of the more commonly used test target metrics are briefly portrayed below; the list is not comprehensive.

1.2.2 SHARPNESS AND RESOLUTION

The camera sharpness can be measured using the spatial frequency response (SFR) from the low-contrast edge of a known test target. Since the test target has known properties, an accurate reproduction of the camera system does not filter high frequencies or add energy to the edges. For example, a sinusoidal Siemens star test target can be used for measuring sharpness (ISO, 2017a; Loebich, Wueller, Klingen, & Jaeger, 2007). The texture of the star becomes successively increasingly more closely spaced when approaching the center, making the contrast of the output decrease and undergo other changes. An edge, initially consisting of black and white sides becomes dark gray and light gray on its sides. The modulation transfer function (MTF) can be measured from the star pattern along the radii of a circle for a range of different angles. There are two important values to observe in the MTF curve: the mid spatial frequency MTF 0.50 information about the local contrast and how sharp the edges appear and the high spatial frequency MTF 0.25 (or MTF 10) information on fine detail sharpness (Figure 1). As an optical system, an MTF-equivalent measure can also be attained from the human eye (Campbell & Green, 1965).

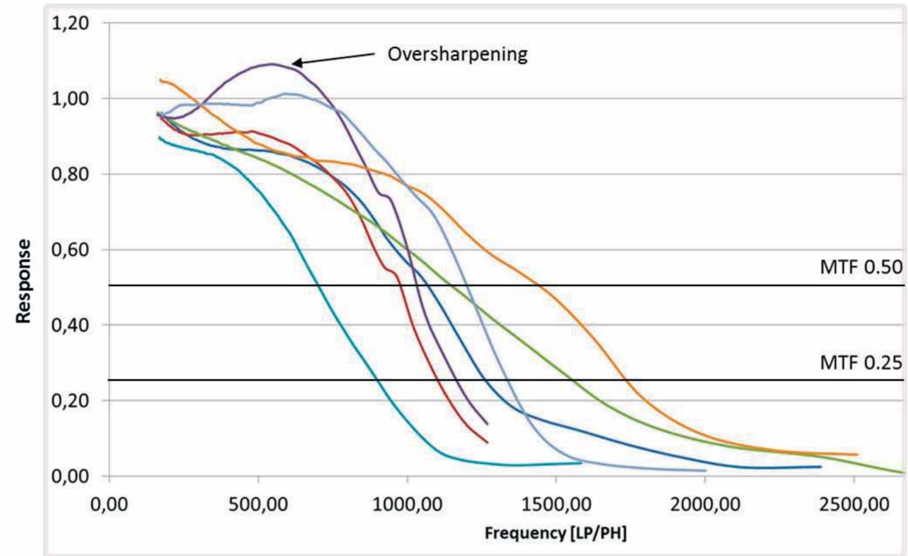

Figure 1 MTF measurements for 7 different imaging systems plotted against line pairs/picture height (LP/PH). There are two important values to observe in the MTF curve: the mid spatial frequency MTF 0.50 information about local contrast and how sharp the edges appear and the high spatial frequency MTF 0.25 information about the fine detail sharpness. The curves that go above 1.00 result from an oversharpening algorithm that increases edge contrast by darkening and lightening the areas near edges. Although it makes the image look sharper, this enhancement can also backfire as it might also make the image look unnatural.

1.2.3 NOISE

The SNR measure is defined as the ratio of the average signal value (P_{signal}) to the standard deviation of the signal value (P_{noise}).

(4) $$SNR = 10log\frac{P_{signal}}{P_{noise}}$$

SNR does not consider that the human visual system (HVS) reacts differently to the spatial distribution of noise and recognizes chroma noise differently the luminance noise. Therefore, a visual noise measure has been developed that can better quantify how well a human observer would recognize noise (IEEE, 2017; ISO, 2017b; Wueller, Matsui, & Katoh, 2019).

1.2.4 OPTICAL DISTORTIONS

Geometrical distortion can be defined as the percentage of the nominal height by which the corners of an image are offset from their ideal location. The two main types of geometric distortions for optical systems are barrel distortion (negative) and pincushion distortion (positive). Uniformity deviations, such as luminance shading and vignetting, are due to the angle at which the light hits the aperture. When light passes directly from the center of the image, the aperture appears round; however, on the sides where light passes from an angle, the aperture becomes elliptical, reducing the amount of light being captured. Another distortion related to the optics is lateral chromatic displacement, which can be seen as color fringes (often purple, blue or red) around high-contrast sharp edges in the image. This is caused by lens refraction being dependent on the wavelength. Different wavelengths of light are being magnified differently by the lens, resulting in them being focused at different positions on the focal plane of the sensor. Each of the above-mentioned distortions can be measured using a test target with black dots over a uniform white background (IEEE, 2017).

1.2.5 COLOR

Colors can be measured by comparing a known property of a test target chart, for example, GretagMacbeth (McCamy, Marcus, & Davidson, 1976), against the output of the system under study. Comparisons are usually made in the CIELAB color space, as it is designed to be perceptually uniform and to account for the spatial-color sensitivity of the human eye (ISO, 2008). The CIELAB color space expresses color as three values: L^* for the lightness from black (0) to white (100), a^* from green (-) to red (+), and b^* from blue (-) to yellow (+). CIELAB was designed such that the same amount of numerical change in these values corresponds to roughly the same amount of visually perceived change. Four common measures are generally used: ΔC = the difference in color chrominance (saturation), ΔH = the difference in color hue, ΔL = the difference in color luminance, and ΔE = the total difference in the $L^*a^*b^*$ space (Figure 2). Another color-related metric is the color uniformity performance of the imaging system. Color uniformity or color shading has two sources. First, the angle at which light strikes the sensor affects how much light is collected by each pixel. Another source is the infrared (IR) filter common in camera modules. When light rays enter the IR filter at an angle, the cutoff wavelength of the filter shifts toward shorter wavelengths. As with luminance shading, where the corners of the image can be darker than the center, color shading is most visible as a color difference in the corners compared to the center of the image, making the colors non-uniform across the imaging plane (I3A, 2007; Wueller, 2006).

- ΔC = difference in color chrominance (saturation)

$\Delta C = C_{ref} - C_{sample}$

- ΔH = difference in color hue (color tone)

- ΔL = difference in color luminance

$\Delta L = L_{ref} - L_{sample}$

- ΔE = difference in L*a*b* color coordinates

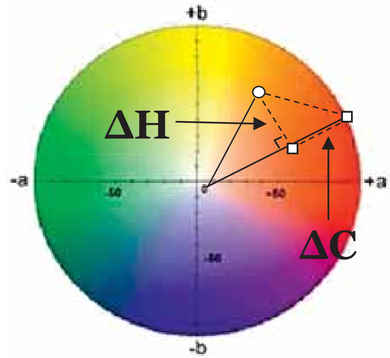

Figure 2 Color metrics comparing the distance of color patches between a reference and the tested image in the CIE L*a*b* space

The benefit of using test chart metrics is their efficiency compared to time-consuming subjective experiments. Unfortunately, objective metrics cannot completely replace subjective testing. For example, the color difference ΔE does not provide information about the preference of colors; it only indicates that the output of the tested imaging system differs from the reference. Using test charts and objective metrics to characterize and measure the properties of imaging devices is not a valuable approach in product development and benchmarking imaging modules. Characterizing the system performance in controlled situations is also crucial for tuning image signal processing (ISP) pipelines. ISP effectively transforms the raw signal from the sensor into a viewable picture. In addition, it controls the three "A"s of the camera: auto-focus, auto-exposure and automatic white balance algorithms. According to (Zhou & Glotzbach, 2007) and (Ramanath, Snyder, Yoo, & Drew, 2005), typical ISP operations are defective pixel correction, noise removal, black level adjustment and color correction. Every camera also introduces different distortion combinations, depending on the camera-specific sensor type, optics and image processing aims.

An interesting industry-driven effort to benchmark camera modules in mobile devices is the Valued Camera eXperience (VCX). The idea is to obtain a comparable estimate of out-of-the-box experience for QoE by using only objective measures. The expert group has agreed and weighed each metric based on their impact on the overall performance of the system. The VCX score is scaled from 0-100, where 100 indicates a device that has the best possible result in every metric achievable using today's technology. This scaling will need to be adjusted as technology develops (Wueller et al., 2018).

1.2.6 IMAGE QUALITY ASSESSMENT ALGORITHMS (IQA)

A popular approach to image quality is to use computational algorithms that estimate and attempt to predict the overall image quality directly from natural images. These IQA algorithms can be divided into three categories: full reference (FR), reduced reference (RR) and no-reference (NR), depending on how much information they require about the original image.

Full reference image quality assessment (FR-IQA) algorithms require a 1:1 pixel level reference image that is used as a yardstick. Some of the test chart metrics can actually be considered simple FR-IQA algorithms. FR-IQA algorithms can be useful, for example, when comparing image file compression methods when the original unpacked image can be used as a reference. Although state-of-the-art FR algorithms manage to achieve very high performances and correlate strongly with human preference evaluations, especially with images degraded with only one distortion type (Li, Bovik, & Wu, 2011), the utility of FR-IQA is somewhat limited in many real world cases where information on the reference is not available.

Reduced reference image quality assessment (RR-IQA) algorithms provide a solution when the reference image is not entirely accessible by extracting features from the reference, which can later be used as additional information when estimating the quality of a distorted image (Cheng & Cheng, 2009; Golestaneh & Karam, 2016; Nuutinen, 2012; Rehman & Wang, 2012). The extracted reference features, also known as side information, can be transmitted efficiently through an ancillary channel. This extra information can be used, for example, to assess and monitor the quality of service of a streamed video signal.

No-reference image quality assessment (NR-IQA) algorithms do not need any information about a reference image and have high real-world application potential. Without any guiding information, other than the image under analysis, the challenges facing NR-IQA are demanding, and their performance has traditionally been mediocre. Therefore, most of the effort has been in developing NR metrics that are based on some assumption that one or more known distortion types, such as blur, white noise, and image compression, have distorted the images (Capodiferro, Jacovitti, & Di Claudio, 2012; Mittal, Soundararajan, & Bovik, 2013; Zhang, Zhang, Mou, & Zhang, 2011; Zhu & Wang, 2012). However, most real-world applications consist of images with multiple concurrent distortions that can have multiple sources, from optical distortions to signal processing and transmission errors. A recent trend in NR-IQA development has been to utilize convolutional neural networks to enhance their performance (Bianco, Celona, Napoletano, & Schettini, 2018; Y. Chen & Jiang, 2018; Kang, Ye, Li, & Doermann, 2014; D. Yang, Peltoketo, & Kämäräinen, 2019; X. Yang, Li, & Liu, 2019).

IQA algorithms can be further divided depending on which features they extract from the images or what type of approach they utilize. The utility of different feature extraction techniques varies depending on the distortion type and use case of the IQA.

Natural scene statistics (NSS) utilize information about the statistical properties of natural scenes (Sheikh & Bovik, 2004). Different image distortions change these statistics compared to the original unaltered image. Some of the statistical properties of natural images, such as the power spectrum distribution, remain valid even across different scene contents (Geisler, 2008).

HVS-based approaches consider the body of knowledge attained with psychophysics about the low-level human visual system processes of the point-spread function (PSF) of the human eye (Campbell & Green, 1965), contrast sensitivity function (CSF) (Campbell & Robson, 1968) and the higher sensitivity to luminance changes than chroma changes in colors (De Valois & De Valois, 1991). Another valuable input from visual perception research to IQA algorithms is masking. For example, certain regions of an image can hide distortions better than other regions, a finding that can be attributed to visual masking (Legge & Foley, 1980).

The structural similarity (SSIM)-based metrics aim to take the texture of the image into account. The premise is that pixels have strong inter-dependencies, especially when they are spatially close, which carry information about the structure of the objects in the visual scene. It also considers some of the properties of the HVS, such as luminance masking, where differences between the reference and test image tend to be less visible in bright regions, and contrast masking, where the differences become less visible in areas where there is significant activity or "texture" in the image. The first metric to utilize this structural similarity was the SSIM (Wang, Bovik, Sheikh, & Simoncelli, 2004). The SSIM compares the reference image and the test images for differences in luminance, contrast, and structural similarity. The SSIM has become a popular method because of its good trade-off between accuracy, simplicity, and efficiency (Rehman & Wang, 2012).

Much of the effort in estimating image quality with IQA has been focused on feature extraction methods. These quality features need to be pooled together to obtain a single quality estimate for the image, as humans do not evaluate images as a set of patches but rather as a whole. IQA algorithms use different spatial feature pooling strategies, such as min, max, mean and percentile, to derive the overall quality estimation. Percentile pooling reflects the significance of highly distorted regions guided by the logic that severe distortions dominate perceived quality. A simple mean pooling, on the other hand, calculates the global image quality by averaging the local patch qualities (Temel & AlRegib, 2015). In addition to spatial feature pooling in images, VQA algorithms often require some form of temporal pooling strategies, for which min, max and mean are commonly used.

An established practice is to validate and test the performance of image and video quality assessment I/VQA algorithms with publicly available image databases (Ciancio et al., 2011b; Horita, Shibata, & Yoshikazu, 2008; Jayaraman, Mittal, Moorthy, & Bovik, 2012; Larson & Chandler, 2010; Le Callet & Autrusseau, 2005; Nuutinen, Virtanen, Vaahteranoksa, et al., 2016;

Ponomarenko et al., 2014, 2009; Sheikh et al., 2006; Virtanen, Nuutinen, Vaahteranoksa, et al., 2015). These databases include sets of images or videos that have undergone some type of distortion and have subjective preference judgment scores attached to them. This enables the training and validation of algorithms for different types of degradation that can have different impacts on the perceived quality of the images or videos. Most of the databases include only images with single distortions; there are few recent databases that incorporate multiply distorted images or videos (Ciancio et al., 2011a; Ghadiyaram & Bovik, 2016; Nuutinen, Virtanen, Vaahteranoksa, et al., 2016; Virtanen, Nuutinen, Vaahteranoksa, et al., 2015). Most of the databases made available by various researchers have been indexed by Qualinet, which has become a valuable resource in the field (Fliegel, 2013).

2. EXPERIMENTS

Each publication is presented briefly below, and detailed descriptions from each study are given in their related chapters.

Publication I describes and introduces an experiment builder software that is specifically designed for subjective image quality measures. The software can be used to conduct many of the standardized subjective methods such as PC, Triplet, ACR, DSIS, DSCR, and SSCQE. It is an essential component of all experimental designs shown in other publications of the thesis. It has also been extensively used in other studies conducted by our research group and partners (Hakola, 2013; Leisti, Radun, Virtanen, Nyman, & Häkkinen, 2014; Nuutinen, Oittinen, & Virtanen, 2012; Nuutinen, Valkonen, Oittinen, & Virtanen, 2013; Nuutinen et al., 2016; Nuutinen, Orenius, Säämänen, & Oittinen, 2010, 2011, 2012; Nuutinen, Virtanen, & Oittinen, 2014; Nuutinen, Virtanen, Vaahteranoksa, et al., 2016; Radun, Leisti, Virtanen, Nyman, & Häkkinen, 2014; Virtanen, Nuutinen, Radun, Leisti, & Häkkinen, 2015; Virtanen, Nuutinen, & Häkkinen, 2019; Virtanen, Nuutinen, Vaahteranoksa, Oittinen, & Häkkinen, 2015).

Publication II proposes a new method for image quality evaluation, where each evaluation made by the observer is calibrated by showing a slideshow of every other image in the test. The absolute category rating with dynamic reference (ACR-DR) method allows observers to consider the quality differences between all the stimuli in the test, thereby reducing variations during their evaluation task.

Publication III presents a new image quality database made available to the research community, where 480 images were evaluated by 188 observers using the ACR-DR method proposed in Publication II. Providing databases of image files along with their quality ratings is essential for IQA algorithm development. Observer preference in the form of subjective ratings is the ground truth that the IQA algorithms strive to predict and can also be used as an input for machine-learning-based algorithms.

Publication IV expands the database to include video material by having 210 observers evaluate 234 video clips. The temporal aspect of the stimuli creates new types of artifacts and degradations for the imaging system, as well as challenges to experimental design and VQA algorithms. Publication IV also illustrates how free descriptions by the observers can be used to add an additional layer of information to the quality evaluations of the videos and aid in the development of empirically based attribute scales such as sharpness.

Publication V formalizes the method of gathering and analyzing the free descriptions presented in Publication IV by aggregating them into attributes with the aid of the FinnWordNet word lexicon. Finding synonyms using an NLP semantic network such as FinnWordNet makes the process visible and repeatable. Publication V also proposes a terminology lexicon in the form of an image quality wheel for images, similar to other sensory experience evaluation tools such as flavor reference wheels (Chen, Rhodes, Crawford, & Hambuchen, 2014; Gawel, Oberholster, & Francis, 2000; Lawless, Hottenstein, & Ellingsworth, 2012; Lawless & Civille, 2013; Meilgaard, Dalgliesh, & Clapperton, 1979; Zarzo & Stanton, 2009). A formalized terminology lexicon can be used to facilitate communication and understanding between professionals in the multidisciplinary field of image quality. In addition, the study examined whether there is a difference in terminology use when evaluating printed photographs or images presented on a display.

Publication VI reviews the interaction between the attributes created using the free descriptions from Publication V and related quality ratings. A regression-based model on the importance of each individual attribute is presented and linked to the quality rating with valence information. Understanding the relationship between image features and their impact on image quality can aid the development of the image and video quality assessment algorithms examined in Publications III and IV.

2.1 PUBLICATION I

> *Nuutinen, M., Virtanen, T., Rummukainen, O. & Häkkinen, J. (2016) VQone MATLAB toolbox: A graphical experiment builder for image and video quality evaluations. Behavior and Research Methods, 48(1).*

Although not published until 2016, the VQone toolbox had been in development since 2009 and has been used extensively in over 60 experiments in our visual cognition research group laboratory and by our industrial partners (Hakola, 2013; Leisti et al., 2014; Nuutinen, Oittinen, et al., 2012; Nuutinen, Orenius, et al., 2012; Nuutinen et al., 2013, 2014; Nuutinen, Virtanen, Leisti, et al., 2016; Nuutinen, Virtanen, Vaahteranoksa, et al., 2016; Nuutinen et al., 2011a, 2011b; Radun et al., 2014; Virtanen et al., 2019; Virtanen, Nuutinen, Radun, et al., 2015; Virtanen, Nuutinen, Vaahteranoksa, et al., 2015).

2.1.1 INCLUDED STANDARD METHODS

The toolbox can be used to implement a wide range of standardized methods of subjective image and video quality evaluation: ACR, ACR-HR, DCR/DSIS, DSCR, SSCQE, PC and triplet comparison (ISO, 2005a, 2005b, 2005c; ITU, 2008b, 2012b). More importantly, the toolbox enables greater freedom in method development and the exploration of experimental design.

2.1.2 FEATURES

The toolbox contains a question building unit (QBU) that provides an open canvas to position sliders, check boxes, text fields and multiple choice buttons to whatever layout is best suited for the specific experiment purposes (Figure 3).

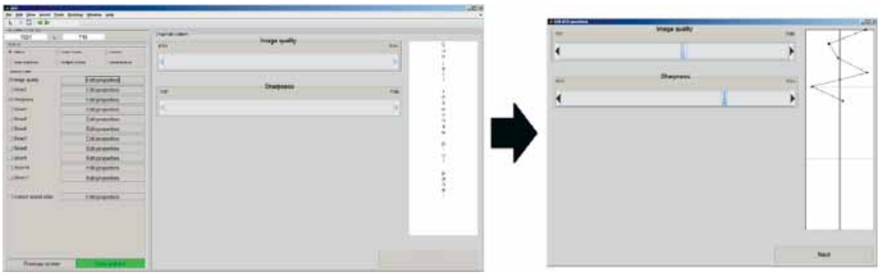

Figure 3 The left side shows the QBU, where sliders, text input fields, check boxes, etc. can be freely positioned on the area. The right side shows the graphical user interface (GUI) provided for the observers during experiments.

There is an option to give the sliders a random starting position, as our experience has shown that the starting position of the slider might introduce a slight bias on how it is used. When the starting position is random, the observers do not mistake the starting position as any type of suggestion of the preferred answer. In addition, with a random starting position, the effort of moving the slider to their preferred position is not constant. For example, a starting position in the center might skew the evaluations toward the middle, while a starting position at the highest possible value would skew them toward higher scores.

A line graph was added on the right side of the GUI to record the previous answers observers have given (see Figure 3 b). The graph's purpose is to help them remember the previous answers and encourage the use of the whole length of the given scale. The line graph is separated in sections for each content so that evaluations related to previous scenes would not confuse the observer when they start to evaluate another scene. The dynamic reference absolute category rating (ACR-DR) method presented in Publication II (Nuutinen, Virtanen, Leisti, et al., 2016) was a significant part of the VQone software development. The new ACR-DR method was later used for the CID2013 image database presented in Publication III (Virtanen, Nuutinen, Vaahteranoksa, et al., 2015). The freedom to design the layout of the GUI in the QBU and add text boxes and graphical scales allowed the use of mixed methods such as interpretation-based quality (IBQ) (Radun, Virtanen, Nyman, & Olives, 2006), where the quantitative evaluations and ratings of the stimuli are combined with observers' free descriptions explaining on which elements their rating were based upon. This method was crucial for the studies presented in Publication V and VI.

2.2 PUBLICATION II

Nuutinen, M., Virtanen, T., Leisti, T., Mustonen, T., Radun, J., & Häkkinen, J. (2016). A new method for evaluating the subjective image quality of photographs: Dynamic Reference. Multimedia Tools and Applications, 75(4).

The dynamic reference (ACR-DR) method was developed for subjective image quality experiments in which original or undistorted images are unavailable. Such situations can occur, for example, in benchmarking studies of imaging systems. Another issue that the method strives to address is the dilemma in the use of reference images in tasks involving preference opinions. With multiple distorted images, selecting a single good or bad reference image can weigh the evaluations toward properties that are most visible on that reference. Without any reference, on the other hand, the observers will create their own 'internal' standards for image quality that are unknown to the researcher, which causes higher variance and numerous outliers in the data. Without a way to anchor the results, the results cannot be compared with earlier or later experiments with different samples without some type of realignment study.

2.2.1 ACR-DR METHOD

In ACR-DR, the observer sees a randomized slideshow of test images with corresponding content, i.e., the dynamic reference, prior to the evaluation task (Figure 4). As the observer views the other test images in the slide show, the observer forms a general gist of the overall quality variation within the set of test images. In this respect, the ACR-DR method partly resembles the SAMVIQ method, which offers access to several samples of a video sequence and the freedom to review and modify their ratings as desired (ITU, 2007; Mantiuk et al., 2012).

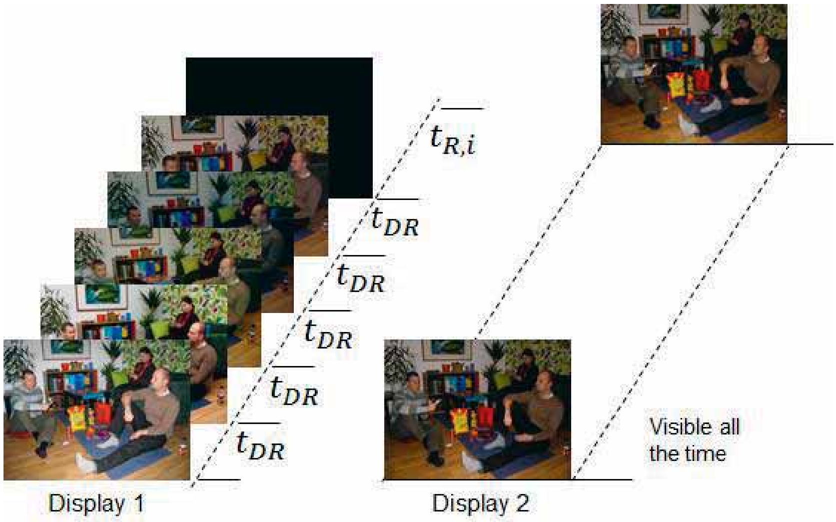

Figure 4 The ACR-DR method presents a reference image set on one display and the test image on the other display. t_{DR} is the display time of one image in a set of reference images.

2.2.2 EXPERIMENTAL SETUP

The ACR-DR method can be seen as a compromise between the ease of use and efficiency of the ACR (ITU, 2008a) and the discriminability of PC (ITU, 2012b). Our hypothesis was that the dynamic reference image set improves the discriminability in determining differences among test images. The drawback of the dynamic reference image set is the increased test duration compared to the ACR method. Three viewing times, 0.25, 0.50 and 0.75 s, for the reference image sets were investigated. These times were selected from a pre-test, where the longest time of 0.75 s was still shorter than the PC test time with 10 images. A total of 75 observers (55 % female) participated in the experiment and were separated into five groups of 15 observers each. The PC was the forced choice 2-AFC method described earlier, and the ACR and ACR-DR used a graphical continuous slider (0 – 100) to evaluate the general quality of the test images. A continuous quality scale was favored over category ranking because categorized attributes can introduce bias (Teunissen, 1996).

Table 2. *Table 2: Overview of the five studies. Testing methods, test durations and number of observers.*

Method	Test duration average	Test duration st. dev.	Observers
ACR-DR25	15,2	3,1	9 women, 6 men
ACR-DR50	18,6	2,9	8 women, 7 men
ACR-DR75	23,2	4,3	8 women, 7 men
ACR	10,2	1,7	8 women, 7 men
PC	26,4	5,3	8 women, 7 men

2.2.3 DISCUSSION

As the PC method has a very high discriminatory power(Gescheider, 1985; ITU, 2008a; Mantiuk et al., 2012), we could assume that it can identify differences among images (if present). We estimated the accuracy of the ACR and ACR-DR methods as a percentage of the discriminatory power compared to the PC method. On average, the PC method differentiated 39 of 45 image pairs. Thus, the accuracy of the ACR method was 60.0 % (23/39 image pairs), the accuracy of the DR25 and DR50 methods was 74.4 % (29/39 image pairs), and the accuracy of the DR75 method was 76.9 % (30/39 image pairs). These results show that the ACR-DR method has higher discriminatory power than the single stimulus ACR. The downside of this is that ACR-DR takes longer to conduct than the ACR method.

2.3 PUBLICATION III

Virtanen, T., Nuutinen, M., Vaahteranoksa, M., Oittinen, P. & Häkkinen, J., (2015). CID2013: a database for evaluating no-reference image quality assessment algorithms. IEEE Transactions on Image Processing, 24(1).

An established practice is to validate and test the performance of a new image quality algorithm with publicly available image databases such as LIVE, TID2008, TID2013, IVC, MICT, LIVE(MDIG), BID, and CSIQ (Ciancio et al., 2011b; Horita et al., 2008; Jayaraman et al., 2012; Larson & Chandler, 2010; Le Callet & Autrusseau, 2005; Ponomarenko et al., 2014, 2009; Sheikh et al., 2006). This study presents a new image database, the CID2013- Camera Image Database, consisting of 480 images gathered from six studies with 188 observers. In contrast to previous image databases, this database uses retail cameras instead of introducing distortions via post-processing. Retail cameras contain images that can have enhancements and distortions that are multidimensional and more subtle in nature and thus constitute a more challenging and ecologically valid database to train and validate image quality algorithms. A comparison to eight other commonly used databases is given in Table 3.

Table 3. *Table 3: Comparison of the CID2013 database with comparable publicly available databases.*

Database	IVC(I)	LIVE(I)	MICT	TID2008	TID2013	CSIQ	BID	LIVE (MDIQ)	CID2013
Year	2005	2006	2008	2009	2013	2009	2011	2012	2013
Images	195	1011	196	1725	3025	930	585	405	480
Rated	185	779	196	1700	3000	866	585	405	480
Contents	10	29	14	25	25	30	585	15	8
Distortion types	4	5	2	17	24	6	5 cat.	3	12–14
Distortion levels	5	5-9	6	4	5	4-5	N/A	16	N/A
Cameras	N/A	KODAK-CD	N/A	KODAK CD+1	KODAK CD+1	N/A	N/A	N/A	79
Simultaneous distortions	1	1	1	1	1-2	1	multiple	2	multiple
Avg. no. of ratings per image	15	23	16	33	9	5-7	11	18–19	31
Method	DSIS	ACR-HR	ACR	PC	PC	Custom	ACR	ACR-HR	ACR-DR
Data	DMOS	DMOS	RAW	MOS+δ	MOS+δ	DMOS+δ	RAW	DMOS+δ	RAW
Scale	1-5	0-100	1-5	0-9	0-9	0-1	0-5	0-100	0-100
Viewing distance	6 Hs	2-2.5 Hs	4 Hp	varying	varying	70 cm	N/A	4 Hs	~80 cm
Screen	CRT	21" CRT	17" CRT	19" LCD & CRT + online	lab & online	24" LCD	17" CRT	LCD	24" LCD
Image resolution	512 x 512	~768 x 512	768 x 512	512 x 384	512 x 384	512 x 512	varied	1280 x 720	1600 x 1200
Display Gamut	N/A	N/A	N/A	varying	varying	sRGB	N/A	N/A	sRGB
Laboratory illumination	N/A	N/A	Low	varying	varying	N/A	N/A	"normal indoor illumination"	5800 K, 20 lx ambient
Format	BMP	BMP	BMP	BMP	BMP	PNG	JPG	BMP	JPG
Subjects	15	20-29	16	838		25	180	37	188
Expert / Naive	expert	naive	naive	naive	naive	N/A	naive	naive	naive
Vision testing	N/A	confirmed verbally	N/A	No	No	N/A	N/A	confirmed verbally	tested
Age	N/A	students	students	N/A	N/A	21–35	N/A	23 - 30	18–44
Female	N/A	minority	N/A	N/A	N/A	N/A	N/A	minority	67 %

DSIS = Double-Stimulus Impairment Scale, ACR = Absolute Category Rating, HR = Hidden Reference, DR = Dynamic Reference, PC = Paired Comparison | Hp = Picture height, Hs = Screen height | δ = Standard deviation | N/A = Not applicable / Not available

2.3.1 IMAGE PROCESSING

CID2013 includes images that have been captured using 79 different cameras or ISP pipelines. ISP transforms the raw signal from the sensor to a viewable picture. In addition, ISP controls the three "A"s of the camera: auto-focus, auto-exposure and automatic white balance algorithms. The variation in ISP quality comes from the choices and differences in signal processing thresholds. For example, to obtain a better exposure on a dark scene, one needs to increase the sensitivity of the sensor; however, this increases noise. How much noise is allowed before the de-noising algorithm starts to reduce it? ISP is always a compromise between computing power, batter consumption and image quality. Using ISP represents actual photographs that would be produced by different cameras with identical lens and sensor characteristics. The cameras used in CID2013 range from low quality to high quality and include low-, moderate- and high-quality mobile phone cameras; moderate-quality compact cameras; and low- to moderate-quality digital single-lens reflex (DSLR) cameras.

2.3.2 SCENES

The image contents were inspired by the "photospace" approach defined by I3A (I3A, 2007). The photospace statistically describes the picture-taking frequency as a function of the subject illumination level *L* and the subject-to-camera distance *D*. The photospace is defined as a probability distribution of "the probability that an image is taken within a certain limit of subject illumination level *L* and within a certain range of subject-camera distance *D*" (I3A, 2007; Segur, 2000). The images in CID2013 represent the utilization of the photospace, which describes where the camera users take the photographs (Table 4). The selected scenes are representative examples of the most prominent clusters from the photospace (I3A, 2007).

Table 4. *Table 4: The luminance, shooting distance and scene descriptions for the images in CID201*

Cluster	Subject luminance (lux)	Subject-camera distance (m)	Scene description	Example images	Image set	Description
1	2	0.5	Close-up in dark lighting conditions		I-VI	Bar and restaurant setting
2	100	1.5	Close-up in typical indoor lighting conditions		I-VI	Living room environment, indoor portrait
3	10	4.0	Small group in dim lighting conditions		I-VI	Living room environment, group picture
4	1000	1.5	Studio image		I-IV	Studio image generally used in image quality testing
5	> 3400	3.0	Small group in cloudy bright to sunny lighting conditions		I-V	Typical tourist image
6	> 3400	> 50	Landscape image in cloudy bright to sunny lighting conditions		I-VI	Landscape image
7	> 3400	3.0	Small group in cloudy bright to sunny lighting conditions (~3x optical or digital zoom)		VI	General zooming situation
8	> 3400 (outdoors) and < 1000 (indoors)	1.5	Close-up in high dynamic range lighting conditions		V,VI	High dynamic range scene

2.3.3 PROCEDURE

The subjective ratings were made using the dynamic reference ACR-DR method presented in Publication II. However, since the viewing time experiments for Publication II were performed after the experiments in this study, we did not yet have information about the optimal viewing time at hand for the dynamic reference slideshow. Our approximation for the best viewing time was then based on the concept that, on average, eyes fixate three times each second by saccadic eye movements to bring the projection of a local scene region onto the area of the fovea, thereby producing the highest acuity vision (Hollingworth & Henderson, 2002). We ended up assuming that 1 second is sufficient to see locally visible quality artifacts but still is as short as possible to not unnecessarily prolong the dynamic reference image slideshow. To be on the safe side, we also included a 500 ms masking image between the images in the dynamic reference slideshow. The white noise masking image effectively removes the illusion of movement between the images, preventing the attention of the subject from being diverted by differences in the image perspective by clearing the iconic memory buffer in the HVS (Sperling, 1960).

The CID2013 image database consists of six different studies, each providing their own image set, which are then combined with a scale realignment study. In addition to the overall quality rating, we also collected separate sharpness, graininess, lightness and saturation ratings. Image sets I-III differed from sets IV-VI by the scale used in the MOS score. In image sets I-III, the observers were instructed to anchor their evaluations by giving the best score on a continuous graphical scale to the best image in the image scene cluster and a score of 0 to the worst image in the image scene cluster. The idea was to use these as anchors to combine the data between studies without a realignment study. Separate experiments can be considered self-contained, and the MOS values cannot be aggregated into one scale without a realignment study (Sheikh et al., 2006; van Dijk et al., 1995).

However, we noticed that observers did not always choose the same images as the best and worst images as hypothesized. Even when the observers preferred different quality aspects in the images, this subjectivity did not translate into unreliability of the data, as the subjects remained consistent with themselves. A hidden repetition image was included in study 2. The within-subject correlation between individual evaluations of the same image was very high ($r = 0.908$, $p < 0.01$), and Cronbach's alpha reliability coefficient, which is a measure of internal consistency, also gave a very good measure of reliability ($\alpha = 0.951$). However, the inconsistency in selecting the best and worst image between observers prevented us from combining the separate image set data into one dataset without a separate realignment study. Therefore, the instruction to anchor the best and worst images was omitted from the instructions for image set studies IV-VI.

2.3.4 REALIGNMENT STUDY

To fit as many images to the scale realignment study, a single-stimulus ACR method was chosen for its time efficiency. A total of 34 observers (85 % female) rated 112 images using a continuous graphical scale. The images were selected to roughly represent the overall scope of image quality variations from each image set study and cluster combination so that each cluster included 14 images. The subjects were provided a training session, where they evaluated

24 images selected to represent the overall set of images in the scale alignment experiment.

To realign the CID2013 scale, MOS values in CID2013 were transformed into Z scores as described in (Sheikh et al., 2006). The averaged Z scores for each image were then compared against averaged image set-specific MOS scores of 112 images that were part of the scale re-alignment study; see Figure 5.

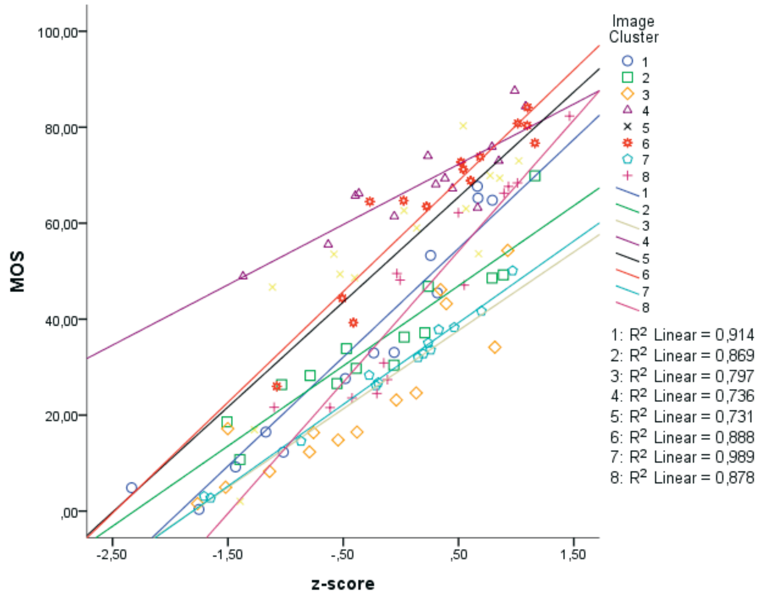

Figure 5 Scatterplot and linear regressions from the realigned study between MOS and z-scores.

2.3.5 IQA PERFORMANCE AGAINST CID2013 DATABASE

IQA algorithms can be divided into three categories, full reference (FR), reduced reference (RR) and no-reference (NR), depending on how much information they require about the original image. The CID2013 database was designed to test NR algorithms, which do not need any information from an unprocessed reference image, thereby using only the information available on the images under evaluation. Table 5 lists the linear correlations with the CID2013 database from the following NR IQA algorithms that were available at the time of publication: BIQI (Moorthy & Bovik, 2010), BRISQUE (Mittal et al., 2012), NIQE (Mittal et al., 2013), BLIINDS-II (Ferzli & Karam, 2009), DESIQUE (Zhang & Chandler, 2013), CPBD (Narvekar & Karam, 2009), FISH (Vu & Chandler, 2012), FISH_bb (Vu & Chandler, 2012), S3 (Vu, Phan, & Chandler, 2012), LPC (Krzic, Donlic, Pejcinovic, & Sersic, 2016), DIIVINE

(Moorthy & Bovik, 2011), Martziliano (Marziliano, Dufaux, Winkler, & Ebrahimi, 2004) and NJQA (Golestaneh & Chandler, 2014).

Before evaluating the performance of an algorithm, a logistic transform to the predicted scores was applied to bring the predicted (objective) and measured (subjective) values onto the same scale and to account for the nonlinear relationships between values (Ma, Lin, Deng, & Ngan, 2012; Sheikh et al., 2006). We used a logistic function with an added linear term (Sheikh et al., 2006):

(5) $$f(x) = \beta_1\left(\frac{1}{2} - \frac{1}{1+\exp(\beta_2(x-\beta_3))}\right) + \beta_4 \cdot x + \beta_5$$

Where β_1, β_2, β_3, β_4 and β_5 are the model parameters chosen to minimize the MSE between the predicted and subjective scores.

Table 5. *Table 5: Linear correlation coefficient after nonlinear regression for realigned MOS.*

Algorithm \ Cluster	1	2	3	4	5	6	7	8	ALL
FISH_bb	**0.69**	0.55	0.31	0.12	0.48	**0.68**	0.79	0.30	**0.49**
FISH	0.65	0.43	0.27	0.05	0.44	0.61	0.73	0.19	0.48
S3	0.67	0.45	0.25	0.17	0.41	0.61	0.79	0.13	0.46
BRISQUE	0.50	0.47	**0.36**	0.40	0.46	0.52	**0.84**	0.27	0.45
BLIINDS-II	0.64	0.45	0.12	**0.54**	**0.52**	0.18	0.57	0.57	0.40
BIQI	0.47	0.30	0.27	0.17	0.36	0.21	0.65	0.26	0.39
DESIQUE	0.57	0.55	0.24	-0.09	0.20	0.35	0.85	0.40	0.34
LPC	0.12	**0.58**	0.28	0.37	0.32	0.46	0.76	**0.66**	0.28
NIQE	-0.47	0.25	0.19	0.34	0.42	0.51	0.78	-0.18	0.26
DIIVINE	-0.33	0.28	0.01	0.19	0.43	0.43	0.64	0.13	0.26
Martziliano	0.44	0.36	0.20	0.11	0.47	0.38	0.53	0.28	0.22
CPBD	-0.43	0.21	-0.03	0.36	0.44	0.36	0.56	0.22	0.19
NJQA	0.10	0.11	0.06	0.25	0.22	0.44	0.57	0.26	0.18

The results show that the performances of FISH_bb (r=0.49), FISH (r=0.48), S3 (r=0.46) and BRISQUE (r=0.56) are comparable with those of the other algorithms. However, the correlations with subjective ratings are low compared to the results of earlier studies with previously published databases (Mittal et al., 2012; Moorthy & Bovik, 2010). For example, the three best-performing algorithms, FISH_bb, FISH and S3, have a Pearson correlation with the LIVE database of 0.944, 0.904 and 0.943, respectively. Clearly, when quality assessment algorithms were developed for images with a single distortion source, their performance leaves significant room for improvement when attempting to predict images with a multi-dimensional concurrent distortion space.

2.4 PUBLICATION IV

Nuutinen, M., Virtanen, T., Vaahteranoksa, M., Vuori, T., Oittinen, P. & Häkkinen, J. (2016). CVD2014 - a database for evaluating no-reference video quality assessment algorithms. IEEE Transactions on Image Processing, 25(7).

As Publication V presented the new image database CID2013, Publication VI presented the CVD2014 camera video database, to be used to train and develop VQA algorithms. As was the case with the CID2013 image database, this database also uses real cameras, with audio, in contrast to previously published databases. Table 6 provides an overview of the following video databases available when Publication VI was published: EFPL-PoliMi (De Simone, Tagliasacchi, Naccari, Tubaro, & Ebrahimi, 2010), ECVQ & EVVQ (Vranješ, Rimac-Drlje, & Grgić, 2013), NYU Video Database (Ou, Xue, & Wang, 2014; Ou, Zhou, & Wang, 2010), NYU Packet Loss Database (Liu, Wang, Boyce, Yang, & Wu, 2009), IRCCyN / IVC Database (Boulos, 2015), LIVE (Seshadrinathan et al., 2010), LIVE mobile (Moorthy et al., 2012), MMSP (SVD) (Lee, De Simone, & Ebrahimi, 2011), CSIQ (Vu & Chandler, 2014), IVP (Zhang, Li, Ma, Wong, & Ngan, 2011), TUM p25 (Keimel, Habigt, Habigt, Rothbucher, & Diepold, 2010), TUM p50 (Keimel, Redl, & Diepold, 2012), AVC HD Database (Staelens, Van Wallendael, Van de Walle, De Turck, & Demeester, 2013), VQEG FR-TV Phase I Database (VQEG, 2000) and the VQEG HDTV database (VQEG, 2010).

Table 6. *Table 6: Comparison of the CVD2014 database with comparable publicly available database.*

Database	EPFL-PoliMi	ECVQ	EVVQ	NYU Video DB	NYU PL DB	IRCCyN / IVC DB	LIVE	LIVE Mobile	MMSP (SVD)	CSIQ	IVP	TUM 1080p25	TUM 1080p50	AVC HD DB	VQEG FR-TV Phase I DB	VQEG HDTV DB	CVD 2014
Year	2009	2013	2013	2008 – 2010	2007	2008 – 2015	2010	2012	2010	2014	2011	2010	2012	2013	2000	2010	2015
Videos	156	90	90	75/68/210	34	varied	150	210	58	228	138	48	20	456	360	740	234
SRC	12	8	8	6/4/6	17	varied	10	10	3	12	10	4	5	8	20	49	5
Clip duration	10 s	12 s	12 s	N/A	20 – 40 s	N/A	~10 s	15 s	10 s	10 s	10 s	10 s	10 s	10 s	8 s	10 s	10 – 25 s
distortion	Transmission error	Compression	Compression	Frame rate, quantization parameter	Transmission error	Compression, Transmission error	Compression, Transmission error	Compression, Transmission error	Spatial & temporal resolution, compression	Compression, Transmission error	Compression, Transmission error	Compression	Compression, viewing scenarios	Transmission error	Compression, Transmission error	Compression, Transmission error	Capture, 3A, signal processing
HRC	12	2	2		9		4	5	various	6	4	3	1	6	16	75	78
video resolution	352 x 288 704 x 576	352 x 288	640 x 480		320 x 240		768 x 432	1280 x 720	1280 x 720	832 x 480	1920 x 1080	1920 x 1080	1920 x 1080	1920 x 1080	1440 x 486 1440 x 576	1920 x 1080	640 x 480 1280 x 720
fps	25 / 30	25	25	Three related databases	12 / 15	Ten databases with various parameters	25 / 50	30	50	varied	25	25	50	25	N/A	N/A	10 – 30
Ratings	34	40	40		15		29	17	16	17-18	42	18	15	24	61-147	24	30
Method	ACR-HR	SAMVIQ	SAMVIQ		SSCQS		ACR-HR	SSCQE + ACR-HR	PC & SSCQS	SAMVIQ	ACR-HR	DSUR	SSMM	ACR	DSCQS	ACR-HR	ACR-HR & IBQ / Scales
Data	RAW	DMOS+δ	DMOS+δ	RAW / MOS+δ	MOS+δ	RAW	DMOS+δ	DMOS+δ	RAW	DMOS+δ	DMOS+δ	RAW	RAW	RAW	DMOS+δ	RAW	RAW

ACR = Single-Stimulus Absolute Category Rating, HR = Hidden Reference, SAMVIQ = Subjective Assessment Methodology for Video Quality, PC = Paired Comparison, SSCQE = Single Stimulus Continuous Quality Evaluation, DSCQE = Double-Stimulus Continuous Quality Evaluation, DSUR = Double Stimulus Unknown Reference | H = Picture height | δ = Standard deviation| N/A = Not applicable / Not available.

Table 6. *Continues.*

Database	EPFL-PoliMi	ECVQ	EVVQ	NYU Video DB	NYU PL DB	IRCCyN / IVC DB	LIVE	LIVE Mobile	MMSP (SVD)	CSIQ	IVP	TUM 1080 p25	TUM 1080 p50	AVC HD DB	VQEG FR-TV Phase I DB	VQEG HDTV DB	CVD2014
Scale	1 – 5	0 – 100	0 – 100	N/A	0 – 100	N/A	0 – 100	0 – 5	0 – 100	0 – 100	1 – 5	0 – 10	0 – 10	1 – 5	1 – 5	1 – 5	0 – 100
Sound	no	no	no	no	no	no	no	no	no	no	no	no	yes	no			yes
Viewing distance	4 – 8 H	5 – 6 H	5 – 6 H	N/A	unrestricted	varied	N/A	unrestricted	2 – 3 H	28"	3 H	3 H	2 H	4 H	5 H	3 H	4 – 6 H
Screen	30" / 19" LCD	19" LCD	19" LCD	N/A	17" LCD	N/A	CRT	4" 10.1" LCD	30"	N/A	65"	24" LCD	23" / 56" / 110”	40" LCD	18" – 20" CRT	24" – 47" LCD	24" LCD
viewing conditions	ITU-R BT.500	ITU-T P.910	ITU-T P.910	N/A	N/A	N/A	N/A	N/A	ITU-R BT.500	N/A	ITU-R BT.500	ITU-R BT.500	ITU-R BT.500	ITU-R BT.500	ITU-R BT.500	ITU-R BT.500	ITU-R BT.500
Subjects	40+40	40	40	16 – 33	57	15 – 30	38	30+17	16	35	42	19	21	40	287	120	210
Expert / Naive	both	naive	naive	N/A	N/A	N/A	N/A	naive	N/A	N/A	both	naive	naive	naive	naive	naive	naive
Vision testing	confirmed verbally	screened	screened	N/A	tested	N/A	no	confirmed verbally	screened	screened	screened	tested	tested	tested	tested	tested	tested
Age	24 – 40	N/A	N/A	N/A	N/A	N/A	N/A	22 – 28	N/A	21 – 32	20 – 38	20 – 30	16 – 27	18 – 34	N/A	N/A	18 – 46
Female	N /A	N/A	N/A	N/A	N/A	N/A	mostly male	mostly male	31 %	N/A	26 %	N/A	9 %	28 %	N/A	N/A	75 %

ACR = Single-Stimulus Absolute Category Rating, HR = Hidden Reference, SAMVIQ = Subjective Assessment Methodology for Video Quality, PC = Paired Comparison, SSCQE = Single Stimulus Continuous Quality Evaluation, DSCQE = Double-Stimulus Continuous Quality Evaluation, DSUR = Double Stimulus Unknown Reference | H = Picture height | δ = Standard deviation| N/A = Not applicable / Not available

2.4.1 VIDEO CAPTURING AND ARTIFACTS

The 234 videos in the CVD2014 database were captured using 78 different cameras (3 DSLR cameras, 4 digital video cameras, 8 digital compact cameras and 63 mobile phone cameras). These different devices create a complex distortion space, as the distortions are related to the video acquisition process rather than being introduced via post-processing. These distortions are very difficult to simulate because they are dependent on not only the optical systems of the capturing devices but also signal processing and sensor characteristics. The raw signal from a sensor includes artifacts such as photon noise, thermal noise, pixel defects, pixel saturation and spatial under-sampling. A low temporal sampling rate results in jerkiness artifacts, which can be perceived as discontinuities in movements. The optics introduce several aberrations such as lens shading and geometrical distortions. The signal control adjusts the 3As of the camera: autofocus (AF), auto-exposure (AE) and automatic white balancing (AWB) algorithms (Nuutinen et al., 2013). A failed exposure or a failed focus induces dark or overexposed video and a loss of detail and sharp edges. Global color errors, such as a green, red or yellow shading in the final video, are often caused by unsuccessful AWB.

The MTF and SNR metrics were measured for characterization of the cameras. The modulation transfer function was measured by the spatial frequency response (SFR)(ISO, 2017a) from the slanted edge area of the MICA test target (Tervonen et al., 2006). The SNR (ISO, 2017b) was measured from the gray patches of the MICA test target, from which the ratio of the average signal value to the standard deviation of the signal value was calculated. The SNR value indicates the noise level as well as noise reduction processing. The MTF value (line pairs per picture height, LP/PH) indicates detailed reproduction and signal sharpening (ISO, 2017a; Koren, 2006; Loebich et al., 2007; Okano, 1997). The IQ-Analyzer software (v. 5.2.7) was used for the analyses.

Figure 6 shows the histograms of the SNR (ISO, 2017b) and MTF (ISO, 2017a) from all cameras for a 1000 lux illumination level. The SNR (ISO, 2017b) was measured from the gray patches of the MICA test target, from which the ratio of the average signal value to the standard deviation of the signal value was calculated. The MTF was measured by the SFR (ISO, 2017a) from the slanted edge area of the MICA test target (Tervonen et al., 2006). From the histograms, it can be observed that the measured values vary, reflecting the varying quality of the cameras. Video sequences with different quality levels are very important if video databases are to be used for the development of VQA algorithms and benchmarking tasks.

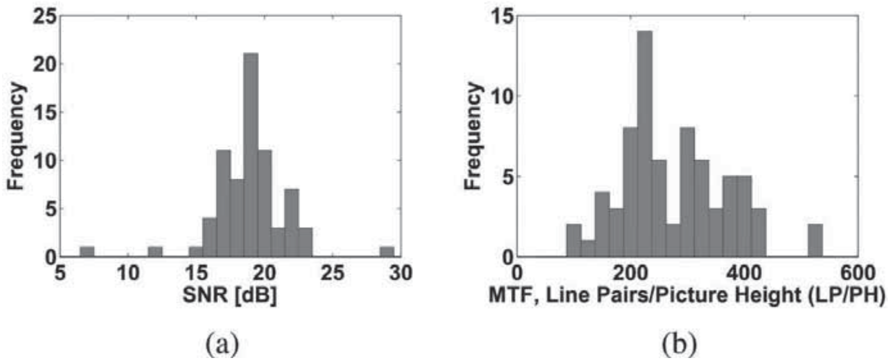

Figure 6 Figure 6: SNR (dB) histograms (a) and MTF50 (LP/PH) histograms (b) for illumination conditions of 1000 lx.

2.4.2 VIDEO SEQUENCES

The challenge of constructing the CVD2014 database was that the video sequences need to be shot by different cameras and still be as similar as possible. In an earlier study, we analyzed 138 randomly chosen videos from an online video sharing service to generate a three-dimensional (distance, illumination, and motion) utilization videospace (Säämänen, Virtanen, & Nyman, 2010). Akin to the photospace approach used with images (I3A, 2007), we used the generated videospace to guide us on which types of video sequences we should generate to make the video database as representative as possible. However, it needed to be possible to replicate the same video sequence multiple times with different devices. We also wanted to make the sequences challenging for the devices to produce distortion artifacts that could be measured by the I/VQA algorithms. The video sequences in the CVD2014 database were captured with five different scenes. The scenes were as follows. **Traffic:** *A bus is driving on a busy road and passes the camera. The camera pans toward the sea, where a man is walking on a walkway.* **City:** *A view from a central location in a city where a man is walking from the outdoors to a tunnel, which includes a gradual change in color temperature and luminance based on the panning camera and moving objects.* **Talking Head:** *The upper body of a man who is talking (in Finnish).* **Newspaper:** *A man is reading a newspaper indoors, and the light changes to light with a different color temperature.* **Television:** *A man is walking to a sofa and picks up an orange from a basket, sits down and switches on a TV, on which a news program begins.* Figure 7 shows three frames from the sequences. The frames are from the beginning, middle and end of the video sequences. The length of the trimmed videos was 10 – 25 s.

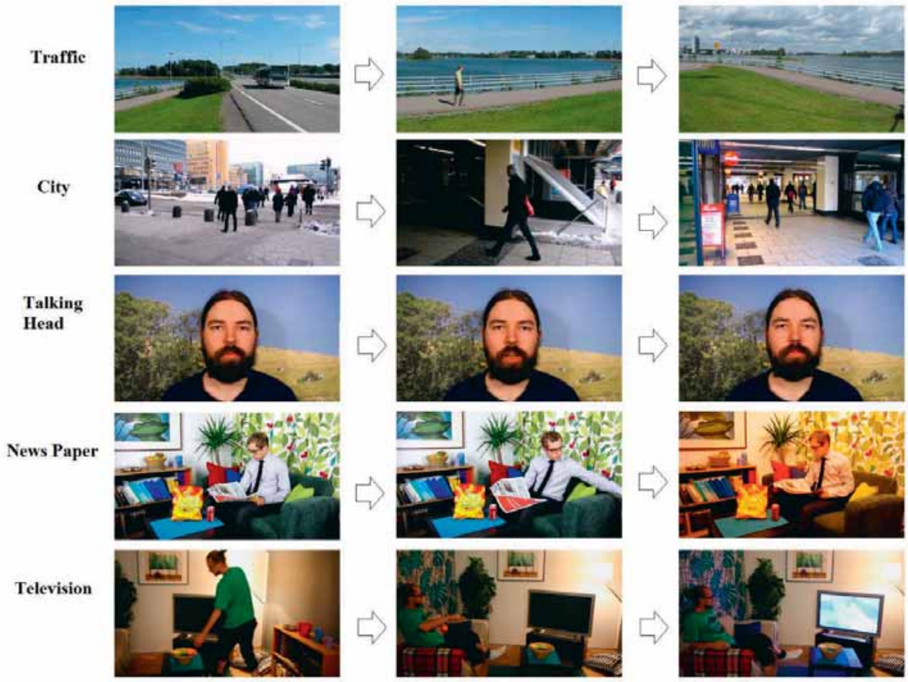

Figure 7 Example frames from the video scenes in the CVD2014 database taken at the beginning, middle and end of the video sequences.

The scenes contain different amounts of spatial and temporal information. Calculation steps according to ITU-T P.910 were followed to create metrics of spatial perceptual information (SI) and temporal perceptual information (TI) for characterizing the level of activity in a video sequence (ITU, 2008a). Instead of using only single SI and TI values to characterize the sequences, we created point clouds to capture the time-series properties of the videos. Figure 8 shows the point cloud values (SI(t),TI(t−1)) of the CVD2014 scenes, where t=2, ... ,T, and T is the number of frames in the video sequence. The values, calculated from the high-quality video sequences, show that the motion and detail levels vary among scenes.

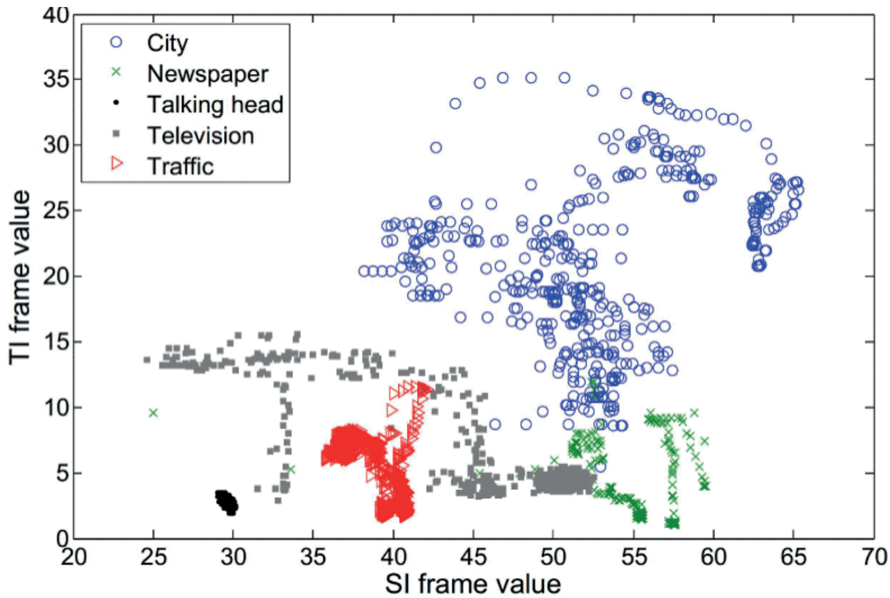

Figure 8 Spatial and temporal activity presented as point cloud values for the example (high-quality) video sequences.

2.4.3 VIDEO POST-PROCESSING

The videos were post-processed to the spatial formats of VGA (640×480 pixels in the CVD I dataset) or HD (1280×720 pixels in the CVD II & CVD III dataset) using the Avisynth script language (v. 2.5) and VirtualDub (v. 1.10.4). The frame rates were maintained at their original values. The Audacity software (v. 2.0.5) was used to normalize the audio volume of the videos, as recommended in ITU-T P.913 (ITU, 2016). The final step was to trim the videos to the same lengths in terms of content so that all sequences began and ended on identical positions in the scene. The videos were compressed using lossless HuffYUV compression with the YUY2 color space and were then deposited into AVI containers. These same post-processed videos were also used for the performance evaluation of the I/VQA algorithms.

2.4.4 PROCEDURE AND VIEWING CONDITIONS

The displays were color calibrated to the sRGB color standard. The luminance level was set to 80 cd/m2, the white point was set to 6500 K, and gamma was set to 2.2. The CVD2014 database is divided into four parts, or sub-datasets: CVD-I, CVD-II, CVD-III and CVD-RA. The CVD-I, CVD-II and CVD-III sub-datasets were constructed from the data from subjective tests 1-6 (TABLE CVD TESTS). Tests 1 and 2 (CVD-I), 3 and 4 (CVD-II) and 5 and 6 (CVD-III) used the single stimulus ACR method with a continuous 0-100 graphical scale (ITU, 2012b).

In tests 1 and 2 (CVD-1), open-ended free descriptions regarding the quality differences between the test videos were also obtained from the observers. These free descriptions were clustered into the attribute classes that define the latent factors of overall video quality. This method (interpretation-based quality, IBQ) of collecting and analyzing free descriptions is described in Publication V and in (Nyman et al., 2006; Radun et al., 2008, 2007; Virtanen et al., 2008). The information gathered from the free descriptions and IBQ method was used to select attribute scales for the remainder of the studies in the CVD database. In addition to the overall quality (Q), the attribute scales of sharpness (S), saturation (Sa), pleasantness of color (PoC), obtrusiveness of change in lighting (OoCiL), lightness (L), motion fluency (MF) and sound quality (SQ) were also gathered. SQ, MF, OoCiL and L were scene-specific scales that were utilized only on scenes for which they were relevant; see Table 7.

Table 7. *Overview of the datasets and test methodologies. The scene numbers are 1: Traffic, 2: City, 3: Talking head, 4: Newspaper, and 5: Television.*

Data sets	Test no.	Scenes	Cameras	No. of videos	Attributes	IBQ	Video resolution	No. observers	Average test time	Age (median)
CVD I	1	1, 2, 3	1 – 9	27	Q	Yes	640x 480	30	1 h 33 min	23
	2	1, 2, 3	10 – 19	30	Q	Yes	640x 480	30	1 h 44 min	24
CVD II	3	2, 3, 4	20 – 32	39	Q, S, Sa and scene-specific attributes (Scene 2: MF, Scene 3: SQ, Scene 4: OoCiL)	No	1280 x720	28	50 min	25
	4	2, 3, 4	33 – 46	42	Q, S, Sa and scene-specific attributes (Scene 2: MF, Scene 3: SQ, Scene 4: OoCiL)	No	1280 x720	33	1 h 6 min	22
CVD III	5	2, 3, 5	47 – 62	48	Q, S, PoC and scene-specific attributes (Scene 2: MF, Scene 3 SQ, Scene 5: L)	No	1280 x720	30	1 h 6 min	24
	6	2, 3, 5	63 – 78	48	Q, S, PoC and scene-specific attributes (Scene 2: MF, Scene 3 SQ, Scene 5: L)	No	1280 x720	32	1 h 1 min	23
CVD RA	7	1 – 5	*	78	Q	No	640x 480, 1280 x720	27	34 min	24

Q = overall quality, S= sharpness, Sa = saturation, PoC = pleasantness of color, OoCiL = obtrusiveness of change in lighting, L = lightness, MF = motion fluency, SQ = sound quality.

2.4.5 REALIGNMENT STUDY

CVD-RA contains data from an additional study in which the mappings from the 18 test-specific quality scales (6 tests × 3 scenes) to the global quality scale were presented. The global scale is valuable when studying and developing VQA algorithms. With the global scale, all samples (234 videos in the case of the CVD2014 database) have the same scale, and the performance analysis of the algorithms can be conducted with a large number of samples.

2.4.6 ANALYSIS OF THE FREE DESCRIPTIONS

To study the descriptive data, open-ended descriptors that depicted the same concepts were aggregated into 17 attribute classes using the procedure described in detail in earlier studies (Nyman et al., 2006; Radun et al., 2008, 2007; Virtanen et al., 2008). For example, the attribute class unsharp included all the descriptors that were related to unsharpness or fuzziness.

PCA was used to extract the main dimensions from the data. The main principal component explained 40 % of the variance in the entire data set. In addition, the combination of dimensions 2 and 3 explained 25 % of the variance, while dimensions 4 to 17 only explained 35 % of the variance (Figure 9). According to this analysis, the subjective overall quality perception of videos can be explained by the sharpness, graininess, color balance, jerkiness and darkness attributes. This result was similar to an earlier study using multidimensional scaling (MDS) to extract the dimensions from free-description attribute data of videos, where the most prominent attributes were sharpness, graininess, jerkiness, faded colors, distorted colors, distorted sound, lip picture-audio synchronization errors, illumination and sound volume (Virtanen et al., 2008).

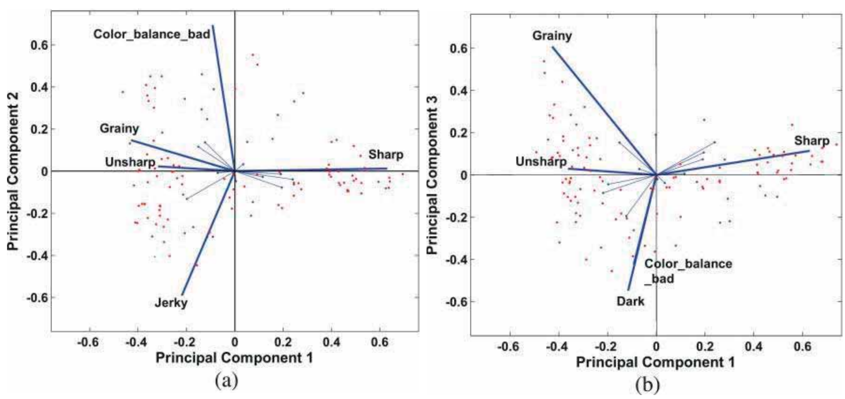

Figure 9 Principal components 1 and 2 (a) and 1 and 3 (b) for the descriptive preferential attribute data.

2.4.7 I/VQA PERFORMANCE AGAINST CVD2014 DATABASE

The performances of several NR-IQA and two VQA algorithms for predicting the video evaluation scores of the CVD2014 database were evaluated. The Video BLIINDS (Saad, Bovik, & Charrier, 2014) and Video CORNIA (Xu, Ye, Liu, & Doermann, 2014) algorithms were the only publicly available NR-VQA algorithms when this study was conducted. Because the number of available NR-VQA algorithms was low, the following NR-IQA algorithms were also selected for the study: BIQI (Moorthy & Bovik, 2010), BRISQUE (Mittal et al., 2012), NIQE (Mittal et al., 2013), DESIQUE (Zhang & Chandler, 2013), FISH (Vu & Chandler, 2012), S3 (Vu et al., 2012), LPC (Krzic et al., 2016), and CPBD (Narvekar & Karam, 2009).

The IQA algorithms extract image quality features from the image and use spatial pooling strategies to obtain a single quality estimate of the image. Because IQA algorithms compute frame-specific scores, these scores also require some temporal pooling into single scalars before comparisons. First, the video sequences were divided into k segments:

(6) $$k = \frac{NoF}{t * fps}$$

Where *NoF* is the number of frames, *fps* is the number of frames per second, and t is the segment duration. The segment-specific values were computed by the temporal pooling operators of *min*, *max* and *mean*. The overall score for the entire video sequence was the average over all segment-specific values. Thus, each IQA algorithm provided three output values.

Before evaluating the performance of an algorithm, a logistic transform to the predicted scores was applied to bring the predicted (objective) and measured (subjective) values onto the same scale and to account for the nonlinear relationships between values (Ma et al., 2012; Sheikh et al., 2006). A non-linear regression of the algorithmic scores using a logistic function suggested VQEG was used to fit the algorithmic scores to the MOS values (VQEG, 2000). This 3-parameter logistic function is written as

(7) $$\hat{Y}(i) = \frac{\beta_1}{1+exp(-\beta_2 * (Y(i) - \beta_3)}$$

Where $Y(i)$ is the quality that is predicted by an algorithm for video i. Non-linear least squares optimization is performed using the Matlab function *nlinfit* (MATLAB R2012a) to find the optimal parameters β that minimize the least square error between the vector of subjective scores MÔS (Equation 7) and the vector of objective scores ($\hat{Y}$).

Table 8 shows the performance of the metrics in terms of Pearson's correlation with the CVD2014 database. Only the best-performing segment operators are shown, e.g., the temporal pooling strategy that provided the highest correlations. In this analysis, the segment duration t was set to 2 s.

According to the results, the BIQI *min* had the highest performance in regard to predicting MÔS values. The second-best algorithm was BRISQUE *min*. Both algorithms were developed to predict overall quality. The third algorithm was FISH_BB *ave*, which was developed to predict sharpness. It is logical that the sharpness algorithm can predict video quality well because according to the analysis from the free descriptions, sharpness is the most important quality dimension when describing the overall quality of the CVD2014 videos.

Table 8. *Pearson's correlation coefficients between Realignment MOS (segment duration = 2 s) and the metric scores after nonlinear regression. Only the best-performing segment operators with the temporal pooling strategy that provided the highest correlations are shown. Bold indicates the best I/VQA performers for each video sequence. The best overall performer was BIQI min, although the second-best-performing method, BRISQUE min, achieved the best performance in 3 out of 5 video sequences.*

Metric	City	Newspaper	Television	Talking head	Traffic	ALL
BIQI *min*	0.602	0.702	0.346	**0.626**	0.416	**0.595**
BRISQUE *min*	**0.726**	**0.768**	0.607	0.484	**0.650**	0.568
FISH_BB *ave*	0.516	0.708	0.730	0.547	-0.030	0.516
LPC *min*	0.497	0.693	0.477	0.596	0.388	0.495
FISH *ave*	0.253	0.724	**0.821**	0.462	0.088	0.437
CBPD *ave*	0.371	0.710	0.436	0.569	0.319	0.390
S3 *max*	0.351	0.602	0.703	0.403	-0.086	0.375
video CORNIA	.0125	0.126	-0.461	0.265	-0.095	0.188
video BLIINDS	-0.032	-0.041	0.103	0.138	0.267	0.122
NIQE *max*	0.019	0.504	0.224	0.285	-0.035	0.090

The performance study revealed that there is room for improvement with regard to I/VQA algorithms when predicting the quality of videos that are captured by different cameras. We believe that the CVD2014 database will provide an important contribution in developing next-generation VQA algorithms capable of predicting the perceived quality of videos captured by different cameras. The presented 17 attribute classes extracted from the free descriptions of the observers could be an interesting starting point for algorithm development.

2.5 PUBLICATION V

Virtanen, T., Nuutinen, M., & Häkkinen, J. (2019). Image quality wheel. Journal of Electronic Imaging, 28(1).

The previous publication briefly discussed how free descriptions and the IBQ method could be used to extract preferential attributes from videos. The ISO defines the preferential attribute as an attribute of image quality that is invariably evident in an image and for which the preferred degree is a matter of opinion, depending upon both the observer and the image content (ISO, 2005a). These preferential attributes are then weighted and summed to create the overall model of image quality (Bech et al., 1996; Engeldrum, 1999, 2004a; Keelan, 2002; Yendrikhovskij, MacDonald, et al., 1999). The concept of the summation of image elements and scene statistics is also used in IQA algorithm development, where various image quality features are extracted computationally from the images and pooled together to create a general estimate of quality (Mittal et al., 2012; Nuutinen, 2012; Temel & AlRegib, 2015).

Perhaps because of its multidisciplinary relevance, the terminology of QoE and image quality is still ill-defined (Augustin et al., 2012). Researchers lack consensus on the most fundamental attributes of image quality and audio-visual quality. For example, usefulness and naturalness were considered defining attributes by Janssen (2001). Sharpness is also thought to be one of the most critical preferential attributes utilized in image quality models (Engeldrum, 1999). Yendrikhovskij et al. (1999) considered naturalness, visibility of details, brightness rendering and chromatic rendering as critical to color television displays.

To facilitate communication and understanding between professionals in approaching the topic of image quality from various fields as well as non-professionals alike, an empirically based image quality attribute terminology lexicon is presented. The multidisciplinary nature of image quality research has a downside of generating discrepancies in terminology, and variable definitions between disciplines can become a problem for mutual comprehension and the sharing of ideas. Reference wheels and terminology lexicons have a long tradition in sensory evaluation fields, such as taste sensory experience studies, where they are used to facilitate communication between interested stakeholders (Chen et al., 2014; Gawel et al., 2000; Kuusinen & Lokki, 2017; Lawless & Civille, 2013; Lawless et al., 2012; Meilgaard et al., 1979; Zarzo & Stanton, 2009).

Pedersen's seminal work can be considered the first attempt to create a standardized lexicon for color print quality. They surveyed attributes from the literature and condensed the results into six dimensions of print image quality, color, lightness, contrast, sharpness, artifacts and physical, which were represented by folded Venn ellipse diagrams (Pedersen, Bonnier, Hardeberg, & Albregtsen, 2010). Our study consists of both printed images and images presented on a display, giving us the possibility to compare how the medium might affect the terminology of the observers.

Contrary to the expert panel or literature review approach of designing terminology lexicons and wheels often used in the sensory evaluation fields (Chen et al., 2014; Gawel et al., 2000; Lawless & Civille, 2013; Lawless et al., 2012; Meilgaard et al., 1979; Zarzo & Stanton, 2009), we opted for an empirical approach based on the observers' free descriptions. Founding the image quality lexicon on empirical data gives us a better understanding of the prevalence distribution between individual attributes. Prevalence can be considered as the visibility and impact of these attributes and how they influence the overall image quality experience (Engeldrum, 2004a). The distribution of different attributes could also indicate to us on which attributes observers base their preferential judgment upon and how that might change depending on the level of quality (Nyman et al., 2010).

2.5.1 EXPERIMENTAL SETUP

The experiments were separated into 7 studies. Studies 1-3 were presented as printed photographs, and studies 4-7 were provided on a display (Table 9). The images were shot using three imaging devices of the same model that were passed around to different individuals to gather as many and as different of images as possible. Altogether, 62 different scenes were selected for the studies. The scenes were intended to represent typical photographs that consumers might capture with their camera devices. Six images had animals, 10 images were architecture pictures, 14 images had bright sunlight, 18 images were night or dark images, 4 images included flowers, 10 pictures were group pictures, 21 pictures were indoor images, 14 pictures were landscapes, 41 pictures were outdoor images, 26 images included people, 15 pictures were portraits, 3 images depicted snow and 2 images were close-ups. The raw signal was manipulated using 60 different ISP pipelines. ISP effectively transforms the raw signal from the sensor to a viewable picture. It also controls the three "A"s of the camera: auto-focus, auto-exposure and automatic white balance algorithms. According to (Zhou & Glotzbach, 2007) and (Ramanath et al., 2005), typical ISP operations are defective pixel correction, noise removal, black level adjustment and color correction. This resulted in the images having multiple overlapping manipulations that might even be counteracting each other, e.g., de-noising vs. sharpening, creating rich stimuli for collecting the free descriptions and creating the image quality wheel.

Table 9. *Breakdown of the seven studies.*

Study	1	2	3	4	5	6	7	SUM
Observers	29	28	30	15	15	15	14	146
ISP's	6	8	6	13	9	9	9	60
Medium	print	print	print	display	display	display	display	
Scenes	15	13	15	6	8	8	8	62

2.5.2 PRINT STUDIES 1-3

The print evaluation task used an SS-ACR method adapted for printed images. Observers sorted the printed images in order of image quality and then scored them on a scale from 0 to 10 using a graphical continuous scale (ITU, 2008b). Observers wore cotton gloves to prevent marking of the prints. After ranking the images, observers were instructed to *"Write down free descriptions for* *each image of the reasons behind your judgment. You don't need to use whole sentences."* The instructions were kept as neutral as possible to prevent any leading questions, as this might influence the way the participants looked at the images (Redi, Liu, Zunino, & Heynderickx, 2011). The experiments were conducted in a room covered with medium gray curtains and tablecloths. The A4 (210 x 297 mm) high-quality glossy prints were presented on an area under 6500 K illumination that varied between 500 and 560 lux. The image files were in the sRGB photospace, and the ICC profile of the professional printing company's printer was added to the image file to ensure correct color management so that the test prints looked exactly like the ISPs would determine them to look.

2.5.3 DISPLAY STUDIES 4-7

The studies followed a modified softcopy version of the ISO 20462-2 Triplet comparison method (ISO, 2005b), where observers saw three images (1920 x 1200 px) depicting the same scene on separate calibrated displays. Instead of only ranking the quality order of images from 1 to 3, each image was rated on a graphical 0 to 10 scale to obtain more gradual information on the quality differences. Giving the same score to two images in a triplet was prevented. Observers were instructed to write their free descriptions with the exact same verbal instruction as in the print studies. The calibration values for the displays were 80 cd/m2, 6500 K, and gamma of 2.2. All experiments were conducted in the same laboratory as the print studies. Fluorescent lights (5800 K) were positioned behind the monitors and reflected from the gray curtain to create dim and uniform ambient illumination in the room. The observers' viewing distances (~80 cm, 2 ½ picture heights) were controlled by a line hanging from the ceiling, and they were instructed to keep their forehead steady next to the line.

2.5.4 ANALYSIS OF THE FREE DESCRIPTIONS

The IBQ approach was utilized for gathering the observers' free descriptions from the visual stimuli (Radun et al., 2006). In the IBQ method, the subjects estimate the overall quality of each image and then describe the most distinctive features of its image quality using free descriptions. The IBQ was inspired by sensory profiling methods of other sensory modalities such as taste and touch (Faye et al., 2004; Picard, Dacremont, Valentin, & Giboreau, 2003), and it was first conceived of as a solution to gain more thorough knowledge of user-experience quality in high-quality magazine printing (Nyman, 2002). The methodology has been successfully tested with image quality evaluation (Radun et al., 2010; Radun, Leisti, Nyman, et al., 2008), print quality evaluation (Leisti et al., 2008), video quality evaluation (Radun et al., 2007; Virtanen et al., 2008), stereoscopic quality evaluation (Shibata et al., 2009), and the quality evaluation of 360° videos (Rummukainen, Radun, Virtanen, & Pulkki, 2014).

The observers' free descriptions, e.g., *"very bright, but blurry image",* were aggregated by a two-step process. First, the grammatical nuances and different inflections, e.g., the terms *bright*, *brighter* and *brightest,* were all summed up manually into wider concepts under the term *bright*. Second, the remaining terms were cross-referenced for synonyms, e.g., bright, luminous, and radiant, to form the final attribute: *Bright* (Figure 10). Synonyms were identified using the FinnWordNet version 2.0 lexical database for Finnish, a derivative of the Princeton WordNet. FinnWordNet contains words (nouns, verbs, adjectives and adverbs) grouped by meaning into synonym groups representing concepts. These synonym groups are linked to each other with relations, such as hyponymy and antonymy, creating a semantic network. As the FinnWordNet was created by having the words of the original English (Princeton) WordNet (version 3.0) translated into Finnish by professional translators (Linden & Carlson, 2010), we could use it to also translate the final attributes from Finnish to English. This method is an evolution from the earlier IBQ method, which simply combined the synonyms manually (Radun et al., 2007; Radun et al., 2010; Radun, Leisti, Nyman, et al., 2008; Rummukainen et al., 2014; Shibata et al., 2009; Virtanen et al., 2008).

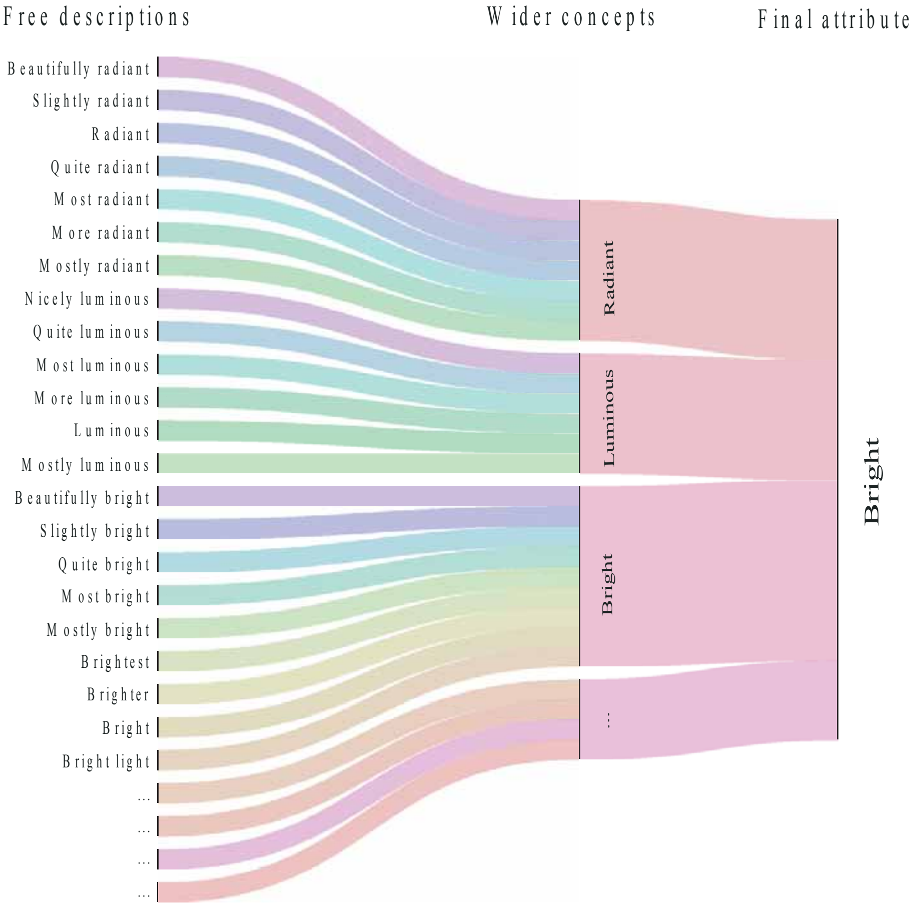

Figure 10 Example of the aggregation and condensing process steps of the free descriptions for creating the final preferential attribute classes.

From the free descriptions of 146 observers, we gathered 39,415 individual quotations. These quotations were then summarized in the first step into 2,742 wider concepts by combining grammatical nuances and different inflections as described above. In the second step, the remaining 2,742 concepts were cross-referenced for synonyms using the FinnWordNet lexical database, making the final count of individual preferential attributes 68, which would represent the empirical basis of the image quality wheel.

The image quality wheel (Figure 11) was inspired by the flavor reference wheels and terminology lexicons from sensory experience fields (Chen et al., 2015; Gawel et al., 2000; Lawless & Civille, 2013; Lawless et al., 2012; Meilgaard et al., 1979; Zarzo & Stanton, 2009). In contrast to the flavor reference wheels, the image quality wheel has an empirical background. With the empirical data, it was possible to add prevalence information to the image quality wheel, where a large attribute frequency is represented by a relatively large area. Colors are used to enhance readability, where each of the free descriptions are given their own hue, which is translated inward to the central core categories. We used a sunburst pie diagram as the model for the image quality wheel, as it provides an effective way to represent the three-tier hierarchy in a condensed manner.

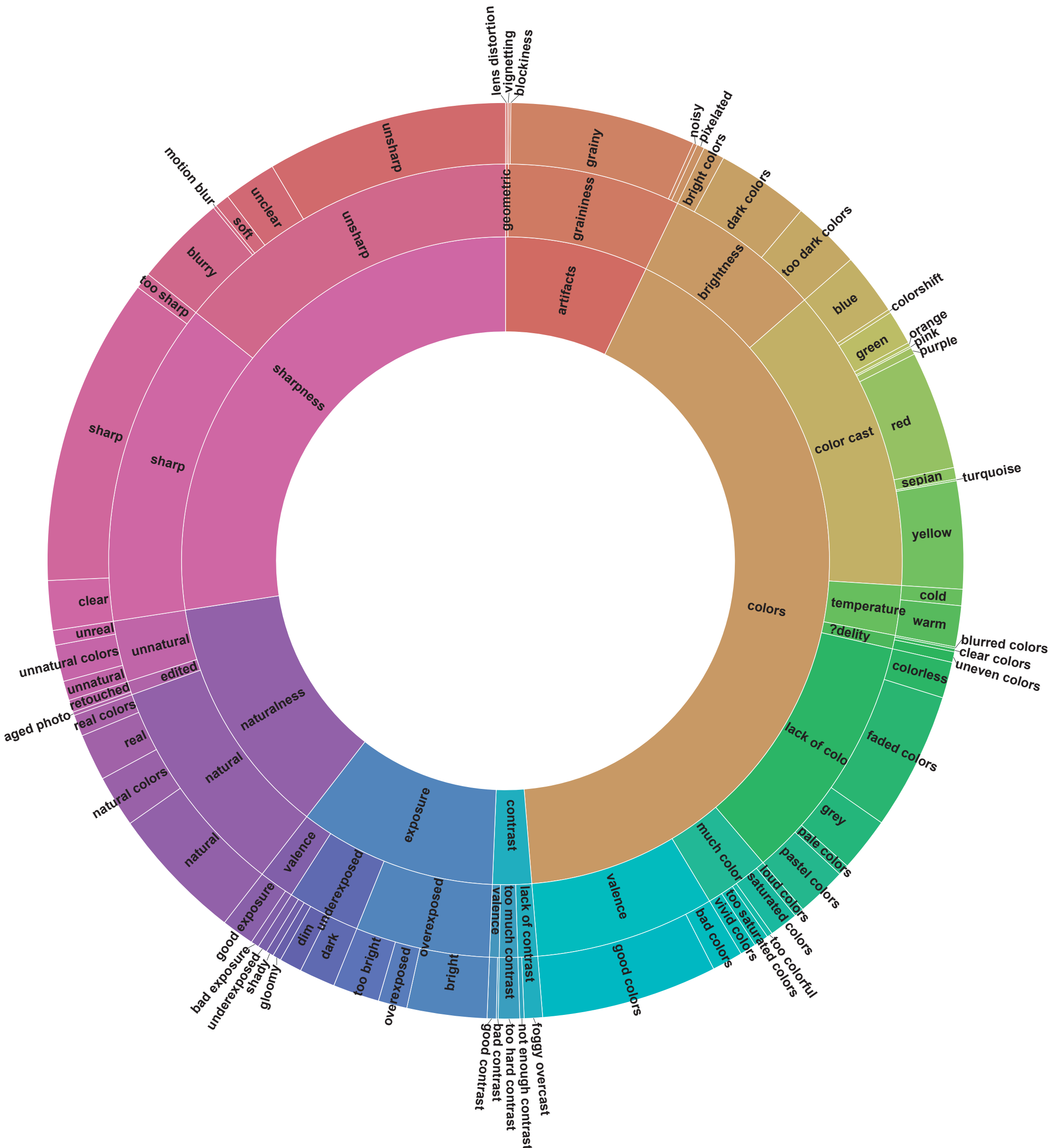

Figure 11 Figure 11: Image quality wheel. All 68 preferential attributes are represented on the outer circle and represented by individual hues. Going inward, the preferential attributes are combined into larger categories. The area of the elements represents the frequency of free descriptions in each category. The larger the area, the more often that preferential attribute has been included in comments by the observers when assessing image quality.

2.5.5 DIFFERENCE IN ATTRIBUTE USE BETWEEN PRINT AND DISPLAY

One of the purposes of this article was to understand whether observers use different terminology when evaluating print images versus images presented on a display. Our results show that the frequency of use of the 68 attributes is highly correlated between printed images and images presented on a display. The Pearson correlation coefficient was r=0.89 for the entire range of attributes. In other words, excluding physical attributes, such as paper gloss, observers base their evaluation on similar image features in both cases. The six core attributes of the image quality wheel, i.e., *artifacts, colors, contrast, exposure, naturalness* and *sharpness*, are almost exactly the same as in Pedersen's model with printed images (Pedersen et al., 2010). Only naturalness has replaced the class related to physical paper properties. What is interesting is that Pedersen ended up with his attributes using a literature review, whereas our attributes are based on naive observers' free descriptions.

2.5.6 DISCUSSION

This study presented an image quality lexicon based on the attributes derived from the free descriptions of 146 participants. It also enhanced the IBQ process of aggregating the preferential attributes from free descriptions by applying NLP techniques. The presented image quality wheel can be used to facilitate communication and understanding between professionals in multidisciplinary fields of image quality.

The image quality wheel provides an efficient way of presenting the hierarchy, variation and prevalence information of preferential attributes of image quality in a single figure. It can be used to facilitate communication and understanding between professionals in the multidisciplinary fields of image quality or as an education tool for observers to help them understand how different attributes might be related on a macro level.

2.6 PUBLICATION VI

Virtanen, T., Nuutinen, M., & Häkkinen, J., Underlying elements of image quality assessment: Preference and terminology for communicating image quality characteristics. Psychology of Aesthetic, Creativity and Art. Psychology of Aesthetic, Creativity, and the Arts. Advance online publication (2020, April 9)

The previous publication presented a lexicon of image quality attributes and studied how observers use terminology in printed images and images presented on a display. This study attempted to explore word use changes between high-quality images and low-quality images and examine the interplay between image quality ratings and word use with multiply distorted images. It builds upon Publication V by creating a regression model to uncover the impact of each of the 68 preferential attributes in the image quality wheel towards image quality rating. The attributes are based on 146 observers giving 39,415 free descriptions while rating the image quality of 62 scenes manipulated by 60 different ISP algorithms. Although our results from Publication V demonstrated that observes use similar terminology when evaluating images presented in print and on displays, we only used the quality ratings from studies 4 to 7, which presented the images on a display, because the experimental methods were different. Print studies used rank ordering akin to absolute category scaling, while the evaluations of the images viewed on displays were collected using a modified triplet comparison method (ISO, 2005b). All experiments were built and conducted using the VQone toolbox presented in Publication I.

2.6.1 EXPERIMENTAL SETUP

A total of 59 (95 % female) observers participated in studies 4-7 in Publication V. None of the participants were professionally involved in photography. All participants had their near visual acuity (ETDRS chart, Precision Vision Inc.), near contrast vision (Near F.A.C.T., Stereo Optical Inc.) and color vision (Farnsworth D-15) tested prior to the experiments. Normal or corrected-to-normal vision was a requirement for participation. The duration of the experiment, which also contained vision tests, instructions, practice and possible breaks, was 1.5 hours.

The studies followed a modified softcopy version of the ISO 20462-2 triplet comparison method (ISO, 2005b), where observers saw three images (1920 x 1200 px) depicting the same scene on separate calibrated displays. Instead of simply ranking the images from 1 to 3, each image was rated on a graphical 0 to 10 scale to obtain more gradual information on the quality differences. Giving the same score to two images in a triplet was prevented, keeping the task as a forced choice comparison. This method gave us both a ranked preference judgment for each triplet and an interval scale evaluation of quality. In addition to the numerical ratings, observers were also instructed to use free descriptions and *"Write down free descriptions for each image of the reasons behind your judgment. You don't need to use whole sentences."* Using as open of instructions as possible, we attempted to not influence the observers in any way, as it has been shown that the instructions can have an impact on the way people look at an image (Radun, Nuutinen, Leisti, & Häkkinen, 2016; Redi et al., 2011).

The triplet comparison method was selected because it forced the observers to make a preference judgment for each triplet on which image they evaluated as the best and which they evaluated as the worst. Combined with the free descriptions, we assumed that this method would provide equal opportunity for positive and negative valence descriptions about the images and not skew the valence distribution because of the evaluation task. For example, with a triplet consisting of three very-low-quality images, one of them still had to be chosen as the best out of the three, and participants had to describe the reason behind that judgment by finding something positive about the image.

2.6.2 RESULTS

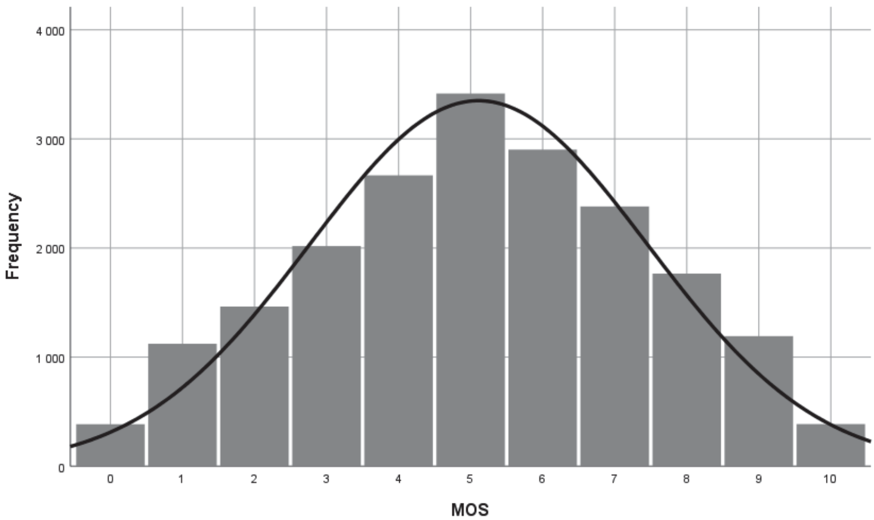

Figure 12 The mean opinion score (MOS) distribution across all four studies follows a normal distribution.

Figure 12 shows that the MOS follows a normal distribution. Multivariate outlier values were identified using the Mahalanobis distance with linear regression analysis. The Mahalanobis distance was compared against a 99.9 % threshold at df = 68 of the chi-squared distribution table. A total of 15.6 % of the values were flagged as multivariate outliers, e.g., having a combination of attributes and MOS that deviates from the overall averages. Further inspection revealed that 91 % of the flagged multivariate outliers were from attributes with frequencies of less than 200 quotations. Some of the multivariate outliers contained all quotations from a single attribute, and removing them would mean that potentially interesting data about the connection between attributes and image quality ratings would be left out of the analysis. In the end, all data were used despite the slight decrease in predictive power for the regression model.

From the 19,692 preference judgments, 5.4 % did not have any verbal description given to them. There were no systematic patterns to be found describing the missing verbal descriptions in the data. A total of 58.0 % of the evaluations had verbal descriptions that could be translated into one attribute, 30.1 % evaluations yielded two attributes per verbal description and 5.4 % gave three attributes. The remaining 1.0 % had more than three attributes, with six attributes being the upper limit.

Each image was ranked either best, worst or in between for each triplet. Figure 13 shows the frequency distribution of each attribute in their corresponding ranks: best out of three (Rank 1), in between (Rank 2) or worst out of three (Rank 3). The attributes were sorted so that those linked most often to the best out of three (Rank 1) are at the top, while attributes linked most often with the worst out of three (Rank 3) are at the bottom.

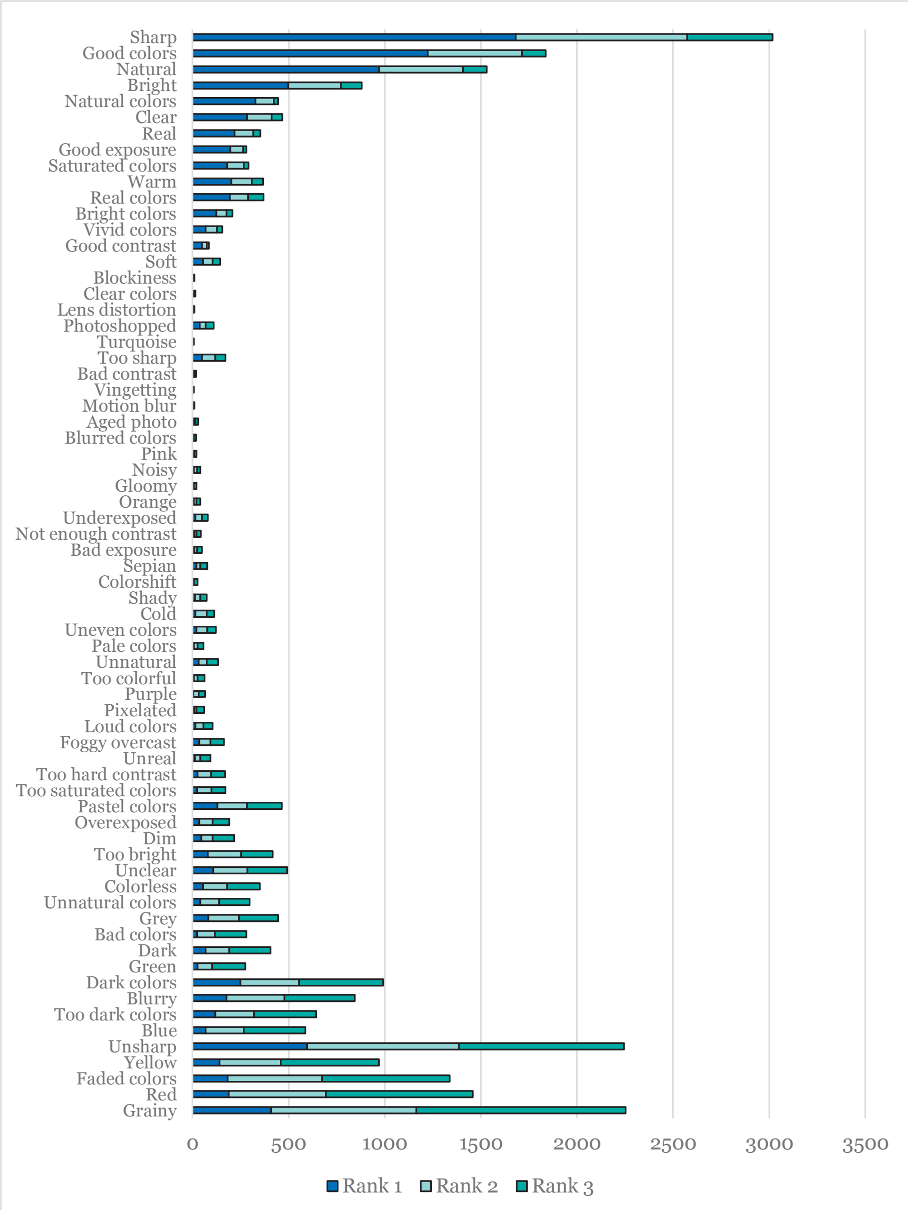

Figure 13 Stacked attribute frequencies separated by ranking of the images in each triplet. The attribute order is sorted such that attributes linked most often to the best out of three (Rank 1) are at the top, while attributes linked most often with the worst out of three (Rank 3) are at the bottom.

2.6.3 IMPACT OF INDIVIDUAL ATTRIBUTES ON PREFERENCE RATINGS

All 68 attributes were entered into the linear regression model as predictors, and MOS was used as the dependent variable ($R^2 = 0.427$, $F(68,19623) = 214.691$, $p < 0.000$), thereby explaining 43 % of the variation in the MOS. A total of 10 attributes, *bad contrast, bad exposure, blockiness, blurred colors, lens distortion, not enough contrast, too colorful, too saturated colors, too sharp and turquoise*, were not significant predictors. To obtain an importance value for each predictor attribute, we used the leave-one-out method, based on the sum of squared errors (SSe), by removing one predictor at a time from the final full model and normalizing the extracted predictor importance to have a cumulative percentage value of 100 % (IBM Corporation, 2017). The higher the importance is, the greater the influence that the attribute has on the predictive model. For example, the attributes *grainy*, *sharp* and *natural* together influence 36 % of the predictive power of the final regression model.

The MOS value for each attribute represents the average image quality of all the images that have been commented to have that attribute. For example, if the image has been described to be *bright,* it would obtain an average image quality score of 7.41; for *unnatural colors,* the image would obtain a score of 4.34. We can also evaluate the valence of each attribute by examining the regression coefficient B of the linear regression model. Negative values decrease the predicted image quality and have a negative valence, whereas positive values increase the quality and have a positive valence. The further the value is from zero, the stronger the effect on the preference evaluations (Table 10).

Table 10. *Attribute frequency, linear regression coefficients, importance and mean opinion scores (MOS). The table is sorted by importance, therein showing the most influential attributes first.*

		Unstandardized Coefficients		Standardized Coefficients					
Predictor	N	B	Std. Error	Beta	t	Sig.	Importance (%)	MOS	Adj. Pred.
(Constant)		5.337	0.029		184.871	0.000			
Grainy	2252	-1.723	0.042	-0.234	-41.222	0.000	14.55	3.67	3.45
Sharp	3017	1.379	0.037	0.212	36.956	0.000	11.70	7.11	6.90
Natural	1532	1.654	0.050	0.189	33.276	0.000	9.48	7.26	7.04
Red	1459	-1.401	0.051	-0.157	-27.416	0.000	6.44	3.88	3.64
Bright	882	1.678	0.063	0.148	26.776	0.000	6.14	7.41	7.23
Good Colors	1838	1.130	0.046	0.14	24.820	0.000	5.28	6.79	6.61
Unsharp	2246	-1.029	0.042	-0.139	-24.613	0.000	5.19	4.41	4.21
Dark Colors	991	-1.318	0.060	-0.123	-22.121	0.000	4.19	3.92	3.74
Blurry	845	-1.330	0.064	-0.115	-20.688	0.000	3.67	3.85	3.66
Real Colors	370	1.880	0.095	0.109	19.890	0.000	3.39	6.56	6.45
Unclear	492	-1.515	0.083	-0.101	-18.319	0.000	2.87	3.72	3.57
Clear	469	1.434	0.084	0.093	17.005	0.000	2.48	6.89	6.79
Yellow	970	-0.984	0.061	-0.091	-16.190	0.000	2.24	4.44	4.26
Bad Colors	281	-1.741	0.108	-0.088	-16.054	0.000	2.21	3.62	3.52
Saturated Colors	291	1.485	0.106	0.076	14.027	0.000	1.68	7.26	6.91
Blue	589	-0.984	0.076	-0.072	-12.902	0.000	1.43	4.21	3.99
Grey	446	-1.075	0.087	-0.068	-12.419	0.000	1.32	4.48	4.18
Dark	405	-1.092	0.091	-0.066	-11.967	0.000	1.23	4.44	4.20
Natural Colors	445	0.979	0.086	0.062	11.318	0.000	1.10	6.49	6.16
Green	274	-1.233	0.109	-0.062	-11.267	0.000	1.09	3.76	3.59
Vivid Colors	154	1.632	0.145	0.061	11.287	0.000	1.09	6.83	6.63
Warm	366	1.070	0.095	0.062	11.300	0.000	1.09	6.52	6.17
Too Dark Colors	644	-0.774	0.073	-0.059	-10.565	0.000	0.96	4.94	4.74
Faded Colors	1338	-0.542	0.053	-0.058	-10.271	0.000	0.90	5.06	4.84
Loud Colors	104	-1.749	0.176	-0.054	-9.959	0.000	0.85	3.52	3.44
Real	352	0.962	0.097	0.054	9.885	0.000	0.84	6.70	6.56
Bright Colors	208	1.120	0.125	0.049	8.957	0.000	0.69	6.62	6.43
Unnatural Colors	297	-0.952	0.110	-0.049	-8.626	0.000	0.64	4.34	4.07
Pixelated	61	-1.562	0.229	-0.037	-6.832	0.000	0.40	2.92	2.66
Colorless	351	-0.661	0.098	-0.037	-6.779	0.000	0.39	4.94	4.56
Uneven Colors	121	-1.093	0.163	-0.036	-6.693	0.000	0.38	4.07	4.00
Dim	216	-0.773	0.123	-0.034	-6.295	0.000	0.34	4.55	4.50
Overexposed	192	-0.812	0.130	-0.034	-6.264	0.000	0.34	4.68	4.55
Too Hard Contrast	170	-0.861	0.138	-0.034	-6.243	0.000	0.33	4.80	4.54
Unnatural	134	-0.906	0.157	-0.032	-5.764	0.000	0.28	4.37	4.07

Table10. *Continues.*

Predictor	N	Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	Standardized Coefficients t	Sig.	Importance (%)	MOS	Adj. Pred.
Purple	66	-1.178	0.221	-0.029	-5.329	0.000	0.24	3.08	2.80
Clear Colors	17	2.189	0.432	0.027	5.062	0.000	0.22	7.18	7.00
Good Exposure	280	0.530	0.108	0.027	4.894	0.000	0.21	6.67	6.31
Unreal	93	-0.936	0.195	-0.027	-4.809	0.000	0.20	3.69	3.55
Foggy Overcast	164	-0.655	0.140	-0.025	-4.667	0.000	0.19	5.07	4.76
Orange	41	-1.264	0.279	-0.025	-4.536	0.000	0.18	4.05	3.75
Sepian	77	-0.896	0.203	-0.024	-4.424	0.000	0.17	4.51	4.35
Noisy	40	-1.171	0.282	-0.022	-4.148	0.000	0.15	3.88	3.77
Cold	113	-0.685	0.169	-0.022	-4.062	0.000	0.14	5.08	4.71
Pale Colors	57	-0.972	0.237	-0.022	-4.108	0.000	0.14	4.40	4.19
Too Bright	418	-0.336	0.089	-0.021	-3.756	0.000	0.12	5.40	5.14
Soft	144	0.509	0.149	0.018	3.406	0.001	0.10	5.97	5.69
Vingetting	8	-2.152	0.630	-0.018	-3.417	0.001	0.10	2.50	2.50
Good Contrast	86	0.571	0.193	0.016	2.960	0.003	0.08	6.48	6.18
Pastel Colors	465	-0.251	0.084	-0.016	-2.969	0.003	0.08	5.31	5.15
Colorshift	28	-0.958	0.337	-0.015	-2.841	0.004	0.07	4.39	4.07
Gloomy	20	-1.081	0.399	-0.015	-2.711	0.007	0.06	4.20	3.90
Aged Photo	31	0.790	0.321	0.013	2.462	0.014	0.05	5.16	5.10
Motion Blur	9	-1.408	0.594	-0.013	-2.373	0.018	0.05	3.67	3.67
Shady	75	0.488	0.206	0.013	2.366	0.018	0.05	6.15	5.84
Photoshopped	111	-0.349	0.172	-0.011	-2.033	0.042	0.04	5.10	4.81
Pink	20	-0.820	0.399	-0.011	-2.058	0.040	0.04	4.30	4.00
Underexposed	81	-0.422	0.201	-0.012	-2.099	0.036	0.04	5.23	4.96
Bad Exposure	50	-0.496	0.258	-0.011	-1.926	0.054	0.03	4.64	4.28
Too Sharp	172	-0.267	0.137	-0.011	-1.947	0.052	0.03	5.43	5.14
Bad Contrast	19	-0.664	0.409	-0.009	-1.623	0.105	0.02	4.58	4.26
Blurred Colors	18	-0.674	0.420	-0.009	-1.605	0.109	0.02	5.39	5.28
Lens Distortion	11	-0.663	0.538	-0.007	-1.233	0.218	0.01	4.91	4.90
Blockiness	9	0.107	0.596	0.001	0.180	0.857	0.00	4.11	4.11
Not Enough Contrast	44	-0.060	0.269	-0.001	-0.225	0.822	0.00	5.30	5.14
Too Colorful	62	-0.143	0.227	-0.003	-0.631	0.528	0.00	5.40	5.14
Too Saturated Colors	172	-0.057	0.138	-0.002	-0.413	0.680	0.00	5.65	5.17
Turquoise	7	-0.427	0.676	-0.003	-0.631	0.528	0.00	3.71	3.71

2.6.4 DISCUSSION

The results from Publication VI show that observers use different terms when describing high quality images and low quality images, supporting the notion made by Nyman et al. (2010) that the subjective decision space can change as a function of preference. The results also suggest that there could be some sort of high-level and low-level distinction to be made in how observers describe image elements, as Leisti et al. (2009) had suggested. The high-level and low-level distinctions can be made mostly from the higher quality images, while low-quality images are mostly only described by more concrete image-fidelity-related elements such as graininess, color casts, lack of sharpness and exposure issues. Certain attributes, such as *brightness*, *naturalness* or *good colors*, seem to be related to high image quality. However, when looking at Table 3, the most important attributes are g*rainy, sharp, natural, bright, red,* and *unsharp*. In addition to the single attribute *natural,* the other attributes seem to be related to image fidelity. This suggests that a certain level of image fidelity has to be achieved before more subjective higher level elements such as naturalness and others can emerge. At least in the case of photographs, processing fluency therefore has an effect on the perception of aesthetic pleasure, as noted elsewhere (Reber et al., 2004). Nevertheless, observers seem to understand that the photographs are meant as representations of the real world, and therefore, images appearing natural will have a significant impact on the perceived quality of the image (Tinio et al., 2011).

Even given the effort to balance out the evaluation task effect on negative or positive bias in word usage, a total of 72 % of the attributes had a negative impact on the preference judgment. Previous studies have had contradictory results on whether the bias is negative (Yendrikhovskij, MacDonald, et al., 1999) or positive (Jacobsen, Buchta, Köhler, & Schröger, 2004). This study differed from previous studies in that we did not interpret the valence of the attribute simply by the interpreted meaning of the words but rather on the effects they had as predictors of the preference judgments. For this study, the overall negative bias can be interpreted in that there are more ways for observers to perceive that an image fails than there are ways to excel. Another explanation could be that observers lack exact words and do not comment when some image-degrading element, such as lack of graininess, is missing.

3. CONCLUSIONS

Image quality is a topic of multidisciplinary relevance, and this work attempted to bridge the disciplines of image engineering and psychology. The work has strived to provide tools and instruments that help drive the field forward and to unify this multidisciplinary research community. Close collaboration with industry partners has given wider perspective to the field of imaging. It has inspired this work toward the applied perspectives in image quality research. Each original publication, although instigated by pure scientific inquiry, has also considered what would be the practical application utility for the industry and the research communities alike. The contributions of this thesis are as follows:

- A purpose-built toolbox for subjective image quality assessment experiments.
- A new subjective image rating method: the ACR-DR.
- Image and video databases to further NR-I/QA algorithm development.
- An image quality preferential attribute lexicon, the image quality wheel, for facilitating communication among researchers.
- Evolution of the IBQ method for analyzing free descriptions using NLP techniques.
- Analysis of the importance between preferential image quality attributes and image quality ratings.

The first publication provided a platform for free-form experimentation with standardized image quality methods. This is the foundation for later works, from which various subjective methods of image quality evaluation could evolve. The second publication focused on the dilemma of using references in subjective experiments by proposing a new method for image quality evaluation: ACR-DR. Inspired by methods, such as ACR, PC and SAMVIQ, ACR-DR allows observers to consider the quality differences of all the stimuli in the test by showing a brief slide show of all the images in the setup, thereby reducing variation during their evaluation task.

The third and fourth original publications set to provide ecologically valid and challenging image and video databases for NR image and VQA (NR-I/VQA) algorithms. State-of-the-art I/VQA algorithms have been mostly trained and tested against databases with only a single distortion applied at a time. The presented databases consist of multiple concurrent distortions that can even mask each other's effects. The image database CID2013 consists of different ISP algorithms that can also include image enhancements and artifacts, such as over sharpening, which were not present in earlier databases. The video database CVD2014 consists of distortions that are related to the video acquisition process instead of introduced degradations from post-processing.

To facilitate communication and understanding among professionals in various fields of image quality as well as non-professionals alike, an attribute lexicon of image quality, the image quality wheel, is presented in the fifth original publication of this thesis. Reference wheels and terminology lexicons have a long tradition in the sensory evaluation fields, such as taste sensory experience studies, where they are used to facilitate communication among interested stakeholders; however, they have not been common in visual experience domains such as image quality. Having consensus on a common terminology can benefit development and research throughout the field.

The final study examined how the free descriptions given by the observers influence the ratings of the images. Understanding how various elements, such as sharpness and naturalness, affect image quality can help one to better understand the decision-making processes behind image quality evaluation. Certain attributes, such as *brightness*, *naturalness* or *good colors*, seem to be related to high image quality. However, the most important attributes are *grainy, sharp, natural, bright, red,* and *unsharp*. In addition to the single attribute *natural,* the other attributes seem to be related to image fidelity. One can hypothesize that a certain level of image fidelity must be achieved before more subjective higher level elements, such as naturalness, can emerge. These results support the concept discussed by Nyman et al. (2010) in that the subjective decision space can change as a function of preference. This information could be an interesting starting point for I/VQA algorithm development, as different sets of rules seem to apply between high quality and low quality.

REFERENCES

Augustin, M. D., Wagemans, J., & Carbon, C.-C. (2012). All is beautiful? Generality vs. specificity of word usage in visual aesthetics. *Acta Psychologica*, *139*(1), 187–201. https://doi.org/10.1016/j.actpsy.2011.10.004

Bech, S., Hamberg, R., Nijenhuis, M., Teunissen, K., Looren de Jong, H. L., Houben, P., … SPIE. (1996). The RaPID Perceptual Image Description Method (RaPID). In B. Rogowitz & J. P. Allebach (Eds.), *Proc. SPIE 2657, Human Vision and Electronic Imaging* (Vol. 2657, pp. 317–328). San Jose, CA, United States: Society of Photo-Optical Instrumentation Engineers (SPIE). https://doi.org/10.1117/12.238728

Berlyne, D. E. (1972). Uniformity in variety: extension to three-element visual patterns and to non-verbal measures. *Canadian Journal of Psychology*, *26*(3), 277–291. https://doi.org/10.1037/h0082436

Bianco, S., Celona, L., Napoletano, P., & Schettini, R. (2018). On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing*, *12*(2), 355–362. https://doi.org/10.1007/s11760-017-1166-8

Boulos, F. (2015). Video Quality Assessment Databases. Institut de Recherche en Communications et Cybernétique de Nantes. Retrieved from http://ivc.univ-nantes.fr/en/pages/view/24/

Burton, M. L., & Nerlove, S. B. (1976). Balanced designs for triads tests: Two examples from English. *Social Science Research*, *5*(3), 247–267. https://doi.org/10.1016/0049-089X(76)90002-8

Campbell, F. W., & Green, D. G. (1965). Optical and retinal factors affecting visual resolution. *The Journal of Physiology*, *181*(3), 576–593. https://doi.org/10.1113/jphysiol.1965.sp007784

Campbell, F. W., & Robson, J. G. (1968). Application of fourier analysis to the visibility of gratings. *The Journal of Physiology*, *197*(3), 551–566. https://doi.org/10.1113/jphysiol.1968.sp008574

Capodiferro, L., Jacovitti, G., & Di Claudio, E. D. (2012). Two-Dimensional Approach to Full-Reference Image Quality Assessment Based on Positional Structural Information. *IEEE Transactions on Image Processing*, *21*(2), 505–516. https://doi.org/10.1109/TIP.2011.2165293

Chen, B., Rhodes, C., Crawford, A., & Hambuchen, L. (2014). Wineinformatics: Applying Data Mining on Wine Sensory Reviews Processed by the Computational Wine Wheel. In *2014 IEEE International Conference on Data Mining Workshop* (Vol. 2015-Janua, pp. 142–149). IEEE. https://doi.org/10.1109/ICDMW.2014.149

Chen, Y., & Jiang, X. (2018). No-reference Image Quality Assessment Based on Convolutional Neural Network. In *2018 IEEE 18th International Conference on Communication Technology (ICCT)* (Vol. 2019-Octob, pp. 1251–1255). IEEE. https://doi.org/10.1109/ICCT.2018.8599897

Cheng, G., & Cheng, L. (2009). Reduced reference image quality assessment based on dual derivative priors. *Electronics Letters*, *45*(18), 937. https://doi.org/10.1049/el.2009.1210

Ciancio, A., da Costa, A. L. N. T. A. L. N. T., Da Silva, E. A. B., Said, A., Samadani, R., & Obrador, P. (2011a). No-Reference Blur Assessment of Digital Pictures Based on Multifeature Classifiers. *Image Processing, IEEE Transactions On*, *20*(1), 64–75. https://doi.org/10.1109/TIP.2010.2053549

Ciancio, A., da Costa, A. L. N. T., Da Silva, E. A. B., Said, A., Samadani, R., & Obrador, P. (2011b). No-Reference Blur Assessment of Digital Pictures Based on Multifeature Classifiers. *Image Processing, IEEE Transactions On*, *20*(1), 64–75. https://doi.org/10.1109/TIP.2010.2053549

Clement, J. (2019). Hours of video uploaded to YouTube every minute as of May 2019. Retrieved November 3, 2019, from https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/

De Simone, F., Tagliasacchi, M., Naccari, M., Tubaro, S., & Ebrahimi, T. (2010). A H.264/AVC video database for the evaluation of quality metrics. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 2430–2433). IEEE. https://doi.org/10.1109/ICASSP.2010.5496296

De Valois, R. L., & De Valois, K. K. (1991). Sensitivity to Color Variations. In *Spatial Vision* (pp. 212–238). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195066579.003.0007

Dodge, S., & Karam, L. (2019). Human and DNN Classification Performance on Images With Quality Distortions. *ACM Transactions on Applied Perception*, *16*(2), 1–17. https://doi.org/10.1145/3306241

Engeldrum, P. G. (1999). Image quality modeling: Where are we? In *(PICS) Image Processing, Image Quality, Image Capture, Systems Conference* (pp. 251–255). IS&T The Society for Imaging Science and Technology.

Engeldrum, P. G. (2000). *Psychometric Scaling: A Toolkit for Imaging Systems Development* (1st ed.). Winchester, MA, USA: Imcotek Press.

Engeldrum, P. G. (2004a). A short image quality model taxonomy. *Journal of Imaging Science and Technology*, *48*(2), 160–165.

Engeldrum, P. G. (2004b). A Theory of Image Quality: The Image Quality Circle. *Journal of Imaging Science and Technology*, *48*(5), 447–457.

Farrell, J. E. (2001). Efficient method for paired comparison. *Journal of Electronic Imaging*, *10*(2), 394. https://doi.org/10.1117/1.1344187

Faye, P., Brémaud, D., Durand Daubin, M., Courcoux, P., Giboreau, A., & Nicod, H. (2004). Perceptive free sorting and verbalization tasks with naive subjects: an alternative to descriptive mappings. *Food Quality and Preference*, *15*(7–8), 781–791. https://doi.org/10.1016/j.foodqual.2004.04.009

Fechner, G. T. (1876). *Vorshule der Aesthetik*. Leipzig: Breitkopf & Härtel.

Fedorovskaya, E. A., de Ridder, H., & Blommaert, F. J. J. (1997). Chroma variations and perceived quality of color images of natural scenes. *Color Research & Application*, *22*(2), 96–110. https://doi.org/10.1002/(SICI)1520-6378(199704)22:2<96::AID-COL5>3.0.CO;2-Z

Ferzli, R., & Karam, L. J. (2009). A No-Reference Objective Image Sharpness Metric Based on the Notion of Just Noticeable Blur (JNB). *IEEE Transactions on Image Processing*, *18*(4), 717–728. https://doi.org/10.1109/TIP.2008.2011760

Fliegel, K. K. (eds. . K. (eds. . K. (2013). QUALINET Multimedia Databases. Retrieved September 29, 2019, from www.qualinet.eu

Gawel, R., Oberholster, A., & Francis, I. L. (2000). A 'Mouth-feel Wheel': terminology for communicating the mouth-feel characteristics of red wine. *Australian Journal of Grape and Wine Research*, *6*(3), 203–207. https://doi.org/10.1111/j.1755-0238.2000.tb00180.x

Geisler, W. S. (2008). Visual Perception and the Statistical Properties of Natural Scenes. *Annual Review of Psychology*, *59*(1), 167–192. https://doi.org/10.1146/annurev.psych.58.110405.085632

Gescheider, G. A. (1985). *Psychophysics: Method, theory, and application* (Vol. 2 nd. edit). 365 Broadway, Hillsdale, New Jersey, USA: Lawrence Erlbaum.

Ghadiyaram, D., & Bovik, A. C. (2016). Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, *25*(1), 372–387. https://doi.org/10.1109/TIP.2015.2500021

Golestaneh, S., & Chandler, D. (2014). No-Reference Quality Assessment of JPEG Images via a Quality Relevance Map. *IEEE Signal Processing Letters*, *21*(2), 155–158. https://doi.org/10.1109/LSP.2013.2296038

Golestaneh, S., & Karam, L. (2016). Reduced-Reference Quality Assessment Based on the Entropy of DWT Coefficients of Locally Weighted Gradient Magnitudes. *IEEE Transactions on Image Processing*, *25*(11), 5293–5303. https://doi.org/10.1109/TIP.2016.2601821

Graf, L. K. M., & Landwehr, J. R. (2015). A Dual-Process Perspective on Fluency-Based Aesthetics. *Personality and Social Psychology Review*, *19*(4), 395–410. https://doi.org/10.1177/1088868315574978

Hakola, V. (2013). *The Impact of Video Sequence Length and Distortion Position in Perceived Quality. MSc. Thesis*. Aalto University.

Heyman, S. (2015, July 29). Photos, Photos Everywhere. *The New York Times*. Retrieved from https://www.nytimes.com/2015/07/23/arts/international/photos-photos-everywhere.html

Hollingworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(1), 113–136. https://doi.org/10.1037/0096-1523.28.1.113

Horita, Y., Shibata, K., & Yoshikazu, K. (2008). MICT Image quality Evaluation Database. Retrieved December 26, 2015, from http://mict.eng.u-toyama.ac.jp/mictdb.html

Hoßfeld, T., Heegaard, P. E., Varela, M., & Möller, S. (2016). QoE beyond the MOS: an in-depth look at QoE via better metrics and their relation to MOS. *Quality and User Experience*, *1*(1), 2. https://doi.org/10.1007/s41233-016-0002-1

I3A. (2007). *CPIQ Initiative Phase 1 White Paper: Fundamentals and review of considered test methods*. Retrieved from http://www.i3a.org/resources/cpiq/

IBM Corporation. (2017). IBM SPSS Modeler 18.1.1 Algorithms Guide. Retrieved October 3, 2019, from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.1.1/en/AlgorithmsGuide.pdf

IEEE. (2017). *IEEE Standard for Camera Phone Image Quality*.

IEEE P2020 Working Group. (2018). *IEEE P2020 Automotive Imaging White Paper IEEE P2020 Automotive Imaging White Paper*.

Instagram Corporation. (2019). Instagram statistics. Retrieved November 3, 2019, from https://instagram-press.com/our-story/

ISO. (2005a). *ISO 20462-1 Photography -- Psychophysical experimental methods for estimating image quality -- Part 1: Overview of psychophysical elements* (Vol. 1).

ISO. (2005b). *ISO 20462-2 - Photography - Psychophysical experimental methods for estimating image quality - Part 2: Triplet comparison method* (Vol. 2).

ISO. (2005c). *ISO 20462-3 Photography -- Psychophysical experimental methods for estimating image quality -- Part 3: Quality ruler method* (Vol. 3).

ISO. (2008). *ISO 11664-4:2008 (CIE S 014-4/E:2007) Colorimetry - Part 4: CIE 1976 L*a*b* Colour space*.

ISO. (2009). *ISO 3664:2009 - Viewing conditions - graphic technology and photography*.

ISO. (2015). *ISO 14524:2009 Photography - Electronic still-picture cameras - Methods for measuring opto-electronic conversion functions (OECFs)*.

ISO. (2016). *ISO 9039:2008 Optics and photonics - Quality evaluation of optical systems - Determination of distortion*.

ISO. (2017a). *ISO 12233:2017 Photography -- Electronic still picture imaging -- Resolution and spatial frequency responses*.

ISO. (2017b). *ISO 15739:2017 Photography - Electronic still-picture imaging -- Noise measurements*.

ISO. (2017c). *ISO 17321-1:2012 Graphic technology and photography - Colour characterisation of digital still cameras (DSCs) - Part 1: Stimuli, metrology and test procedures*.

ISO. (2019). *ISO 12232:2019 Photography - Digital still cameras - Determination of exposure index, ISO speed ratings, standard output sensitivity, and recommended exposure index*.

ITU. (1990). ITU Rep. BT.1082-1: Studies towards the unification of picture assessment methodology. ITU.

ITU. (1997). *ITU-R BT.1128-2 Subjective assessment of conventional television systems*.

ITU. (1998a). *ITU-R BT.1129-2 Subjective assessment of standard definition digital television (SDTV) systems*.

ITU. (1998b). *ITU-R BT.710-4 Subjective assessment methods for image quality in high-definition television*.

ITU. (2007). *ITU-R BT.1788 Methodology for the subjective assessment of video quality in multimedi applications*.

ITU. (2008a). *ITU-T Recommendation P. 910: Subjective Video Quality Assessment Methods for Multimedia Applications* (Vol. 910).

ITU. (2008b). *Recommendation ITU-T P.910: Subjective video quality assessment methods for multimedia applications* (Vol. 910).

ITU. (2012a). *ITU-R BT.2022-2: General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays*.

ITU. (2012b). *ITU-R BT.500-13. Methodology for the subjective assessment of the quality of television pictures* (Vol. 500–13).

ITU. (2016). *ITU-T Rec. P.913 Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribtution quality television in any environment*.

Jacobsen, T., Buchta, K., Köhler, M., & Schröger, E. (2004). The Primacy of Beauty in Judging the Aesthetics of Objects. *Psychological Reports*, *94*(3_suppl), 1253–1260. https://doi.org/10.2466/pr0.94.3c.1253-1260

Janssen, T. (2001). Understanding image quality. In *Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205)* (Vol. 2, p. 7 vols.2-). https://doi.org/10.1109/ICIP.2001.958408

Janssen, T., & Blommaert, F. (1997). Image Quality Semantics. In *Journal of Imaging Science and Technology* (Vol. 41, pp. 555–560).

Jayaraman, D., Mittal, A., Moorthy, A., & Bovik, A. (2012). Objective Image Quality Assessment of Multiply Distorted Images. In *2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)* (pp. 1693–1697). Monterey, CA: IEEE. https://doi.org/10.1109/ACSSC.2012.6489321

Jin, E., & Keelan, B. (2009). Aptina Softcopy Quality Ruler User ' s Manual. Aptina.

Kang, L., Ye, P., Li, Y., & Doermann, D. (2014). Convolutional Neural Networks for No-Reference Image Quality Assessment. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1733–1740). IEEE. https://doi.org/10.1109/CVPR.2014.224

Keelan, B. (2002). *Handbook of image quality: characterization and prediction*. CRC Press.

Keelan, B., & Urabe, H. (2003). ISO 20462: a psychophysical image quality measurement standard. In Y. Miyake & D. R. Rasmussen (Eds.), *Proceedings of SPIE and IS&T, Image Quality and System Performance I* (Vol. 5294, pp. 181–189). San Jose, CA, United States: SPIE and IS&T. https://doi.org/10.1117/12.532064

Keimel, C., Habigt, J., Habigt, T., Rothbucher, M., & Diepold, K. (2010). Visual quality of current coding technologies at high definition IPTV bitrates. In *2010 IEEE International Workshop on Multimedia Signal Processing* (pp. 390–393). IEEE. https://doi.org/10.1109/MMSP.2010.5662052

Keimel, C., Redl, A., & Diepold, K. (2012). The TUM high definition video datasets. In *2012 Fourth International Workshop on Quality of Multimedia Experience* (pp. 97–102). IEEE. https://doi.org/10.1109/QoMEX.2012.6263865

Kokaram, A., Foucu, T., & Hu, Y. (2016). A look into YouTube's video file anatomy. Retrieved November 3, 2019, from https://youtube-eng.googleblog.com/2016/04/a-look-into-youtubes-video-file-anatomy.html

Koren, N. (2006). The Imatest program: comparing cameras with different amounts of sharpening. In N. Sampat, J. M. DiCarlo, & R. A. Martin (Eds.), *SPIE 6069, Digital Photography II* (p. 60690L). San Jose, CA, United States. https://doi.org/10.1117/12.650848

Krzic, A. S., Donlic, M., Pejcinovic, M., & Sersic, D. (2016). Image sharpness assessment based on local phase coherence and LAD criterion. In *2016 International Conference on Systems, Signals and Image Processing (IWSSIP)* (Vol. 22, pp. 1–4). IEEE. https://doi.org/10.1109/IWSSIP.2016.7502724

Kuusinen, A., & Lokki, T. (2017). Wheel of Concert Hall Acoustics. *Acta Acustica United with Acustica*, *103*(2), 185–188. https://doi.org/10.3813/AAA.919046

Larson, E. C., & Chandler, D. M. (2010). Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, *19*(1), 11006–11021. https://doi.org/10.1117/1.3267105

Lawless, L., & Civille, G. (2013). Developing Lexicons: A Review. *Journal of Sensory Studies*, *28*(4), 270–281. https://doi.org/10.1111/joss.12050

Lawless, L., Hottenstein, A., & Ellingsworth, J. (2012). The McCormick spice wheel: A systematic and visual approach to sensory lexicon development. *Journal of Sensory Studies*, *27*(1), 37–47. https://doi.org/10.1111/j.1745-459X.2011.00365.x

Le Callet, P., & Autrusseau, F. (2005, February 2). Subjective quality assessment IRCCyn/IVC database. Retrieved October 3, 2019, from http://www2.irccyn.ec-nantes.fr/ivcdb/

Le Callet, P., Möller, S., & Perkins, A. (2012). *Qualinet White Paper on Definitions of Quality of Experience* (Vol. 1.2). Lausanne, Switzerland. Retrieved from www.qualinet.eu

Leder, H., Belke, B., Oeberst, A., & Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology*, *95*(4), 489–508. https://doi.org/10.1348/0007126042369811

Lee, J., De Simone, F., & Ebrahimi, T. (2011). Subjective Quality Evaluation via Paired Comparison: Application to Scalable Video Coding. *IEEE Transactions on Multimedia*, *13*(5), 882–893. https://doi.org/10.1109/TMM.2011.2157333

Legge, G. E., & Foley, J. M. (1980). Contrast masking in human vision. *Journal of the Optical Society of America*, *70*(12), 1458–1471. https://doi.org/10.1364/josa.70.001458

Leisti, T., Halonen, R., Kokkonen, A., Weckman, H., Mettänen, M., Lensu, L., ... Nyman, G. (2008). Process perspective on image quality evaluation. In S. P. Farnand & F. Gaykema (Eds.), *Proc. SPIE 6808, Image Quality and System Performance V* (p. 68080P). San Jose, CA, United States: Society of Photo-Optical Instrumentation Engineers (SPIE). https://doi.org/10.1117/12.765438

Leisti, T., Radun, J., Virtanen, T., Halonen, R., & Nyman, G. (2009). Subjective experience of image quality: attributes, definitions, and decision making of subjective image quality. In S. P. Farnand & F. Gaykema (Eds.), *Proceedings of SPIE - The International Society for Optical Engineering* (Vol. 7242, p. 72420D). San Jose, CA, United States. https://doi.org/10.1117/12.807142

Leisti, T., Radun, J., Virtanen, T., Nyman, G., & Häkkinen, J. (2014). Concurrent explanations can enhance visual decision making. *Acta Psychologica*, *145*(1), 65–74. https://doi.org/10.1016/j.actpsy.2013.11.001

Li, C., Bovik, A. C., & Wu, X. (2011). Blind Image Quality Assessment Using a General Regression Neural Network. *IEEE Transactions on Neural Networks*, *22*(5), 793–799. https://doi.org/10.1109/TNN.2011.2120620

Linden, K., & Carlson, L. (2010). FinnWordNet – WordNet på finska via översättning. *LexicoNordica*, *17*, 119–140. Retrieved from http://www.ling.helsinki.fi/~klinden/pubs/FinnWordnetInLexicoNordica-en.pdf

Liu, T., Wang, Y., Boyce, J. M., Yang, H., & Wu, Z. (2009). A Novel Video Quality Metric for Low Bit-Rate Video Considering Both Coding and Packet-Loss Artifacts. *IEEE Journal of Selected Topics in Signal Processing*, *3*(2), 280–293. https://doi.org/10.1109/JSTSP.2009.2015069

Loebich, C., Wueller, D., Klingen, B., & Jaeger, A. (2007). Digital camera resolution measurement using sinusoidal Siemens stars. In R. A. Martin, J. M. DiCarlo, & N. Sampat (Eds.), *Proceedings of SPIE - The International Society for Optical Engineering* (Vol. 6502, p. 65020N). https://doi.org/10.1117/12.703817

Ma, L., Lin, W., Deng, C., & Ngan, K. (2012). Image Retargeting Quality Assessment: A Study of Subjective Scores and Objective Metrics. *IEEE Journal of Selected Topics in Signal Processing*, *6*(6), 626–639. https://doi.org/10.1109/JSTSP.2012.2211996

Mantiuk, R. K., Tomaszewska, A., & Mantiuk, R. (2012). Comparison of Four Subjective Methods for Image Quality Assessment. *Computer Graphics Forum*, *31*(8), 2478–2491. https://doi.org/10.1111/j.1467-8659.2012.03188.x

Marziliano, P., Dufaux, F., Winkler, S., & Ebrahimi, T. (2004). Perceptual blur and ringing metrics: application to JPEG2000. *Signal Processing: Image Communication*, *19*(2), 163–172. https://doi.org/10.1016/j.image.2003.08.003

McCamy, C. S., Marcus, H., & Davidson, J. G. (1976). A Color-Rendition Chart. *Journal of Applied Photographic Engineering*, *2*(3), 95–99.

Meilgaard, M. C., Dalgliesh, C. E., & Clapperton, J. F. (1979). Beer Flavour Terminology. *Journal of the Institute of Brewing*, *85*(1), 38–42. https://doi.org/10.1002/j.2050-0416.1979.tb06826.x

Mittal, A., Moorthy, A., & Bovik, A. (2012). No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing*, *21*(12), 4695–4708. https://doi.org/10.1109/TIP.2012.2214050

Mittal, A., Soundararajan, R., & Bovik, A. C. (2013). Making a “Completely Blind” Image Quality Analyzer. *IEEE Signal Processing Letters*, *20*(3), 209–212. https://doi.org/10.1109/LSP.2012.2227726

Moorthy, A., & Bovik, A. (2010). A Two-Step Framework for Constructing Blind Image Quality Indices. *IEEE Signal Processing Letters*, *17*(5), 513–516. https://doi.org/10.1109/LSP.2010.2043888

Moorthy, A., & Bovik, A. (2011). Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality. *IEEE Transactions on Image Processing*, *20*(12), 3350–3364. https://doi.org/10.1109/TIP.2011.2147325

Moorthy, A., Choi, L., Bovik, A., & de Veciana, G. (2012). Video Quality Assessment on Mobile Devices: Subjective, Behavioral and Objective Studies. *IEEE Journal of Selected Topics in Signal Processing*, *6*(6), 652–671. https://doi.org/10.1109/JSTSP.2012.2212417

Narvekar, N. D., & Karam, L. J. (2009). A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection. In *2009 International Workshop on Quality of Multimedia Experience* (pp. 87–91). IEEE. https://doi.org/10.1109/QOMEX.2009.5246972

Nuutinen, M. (2012). *Reduced-reference methods for measuring quality attributes of natural images in imaging systems. Doctoral Dissertation*. Aalto University.

Nuutinen, M, Oittinen, P., & Virtanen, T. (2012). Features for Predicting Quality of Images Captured by Digital Cameras. In *2012 IEEE International Symposium on Multimedia* (pp. 165–168). IEEE. https://doi.org/10.1109/ISM.2012.40

Nuutinen, M., Orenius, O., Säämänen, T., & Oittinen, P. (2012). A framework for measuring sharpness in natural images captured by digital cameras based on reference image and local areas. *EURASIP Journal on Image and Video Processing*, *2012*(1), 8. https://doi.org/10.1186/1687-5281-2012-8

Nuutinen, M., Valkonen, V., Oittinen, P., & Virtanen, T. (2013). Automatic exposure and white balance control in video cameras: Time course characterization and preference. In *2013 8th International Symposium on*

Image and Signal Processing and Analysis (ISPA) (pp. 25–29). IEEE. https://doi.org/10.1109/ISPA.2013.6703709

Nuutinen, M., Virtanen, T., Leisti, T., Mustonen, T., Radun, J., & Häkkinen, J. (2016). A new method for evaluating the subjective image quality of photographs: dynamic reference. *Multimedia Tools and Applications*, *75*(4), 2367–2391. https://doi.org/10.1007/s11042-014-2410-7

Nuutinen, M., Virtanen, T., & Oittinen, P. (2014). Image feature subsets for predicting the quality of consumer camera images and identifying quality dimensions. *Journal of Electronic Imaging*, *23*(6), 061111. https://doi.org/10.1117/1.JEI.23.6.061111

Nuutinen, M., Virtanen, T., Vaahteranoksa, M., Vuori, T., Oittinen, P., & Hakkinen, J. (2016). CVD2014—A Database for Evaluating No-Reference Video Quality Assessment Algorithms. *IEEE Transactions on Image Processing*, *25*(7), 3073–3086. https://doi.org/10.1109/TIP.2016.2562513

Nuutinen, M., Orenius, O., Säämänen, T., & Oittinen, P. (2011a). Potential of face area data for predicting sharpness of natural images. In S. P. Farnand & F. Gaykema (Eds.), *Image Quality and System Performance VIII* (p. 786710). San Francisco Airport, California, USA: IS&T The Society for Imaging Science and Technology / SPIE. https://doi.org/10.1117/12.871996

Nuutinen, M., Orenius, O., Säämänen, T., & Oittinen, P. (2011b). Reference image method for measuring quality of photographs produced by digital cameras. In S. P. Farnand & F. Gaykema (Eds.), *Image Quality and System Performance VIII* (Vol. 7867, p. 78670M). San Francisco Airport, California, USA. https://doi.org/10.1117/12.871999

Nyman, G., Häkkinen, J., Koivisto, E.-M., Leisti, T., Lindroos, P., Orenius, O., ... Vuori, T. (2010). Evaluation of the visual performance of image processing pipes: information value of subjective image attributes. In S. P. Farnand & F. Gaykema (Eds.), *Proceedings of SPIE - The International Society for Optical Engineering* (Vol. 7529, p. 752905). https://doi.org/10.1117/12.839946

Nyman, G., Radun, J., Leisti, T., Oja, J., Ojanen, H., Olives, J.-L., ... Häkkinen, J. (2006). What do users really perceive probing - probing the subjective image quality. In L. C. Cui & Y. Miyake (Eds.), *Image Quality and System Performance III* (Vol. 6059, pp. 605902-605902–605907). https://doi.org/10.1117/12.641612

Nyman, G., Radun, J., Leisti, T., & Vuori, T. (2005). From image fidelity to subjective quality: A hybrid qualitative/ quantitative methodology for measuring subjective image quality for different image contents. *IDW/AD'05 - Proceedings of the 12th International Display Workshops in Conjunction with Asia Display 2005*, (2), 1817–1820.

Nyman, G. (2002). Quality Experience Research: Trying to Understand the Modern Magazine Reader in Multimedia World. London, UK: PPA Professional Publishers Association.

O'Hare, D. P. A., & Gordon, I. E. (1977). Dimensions of the perception of art: verbal scales and similarity judgements. *Scandinavian Journal of Psychology*, *18*(1), 66–70. https://doi.org/10.1111/j.1467-9450.1977.tb00257.x

Okano, Y. (1997). MTF Analysis and its Measurements for Digital Still Camera. *IS&T's 50th Annual Conference*.

Ou, Y.-F., Xue, Y., & Wang, Y. (2014). Q-STAR: A Perceptual Video Quality Model Considering Impact of Spatial, Temporal, and Amplitude

Resolutions. *IEEE Transactions on Image Processing*, *23*(6), 2473–2486. https://doi.org/10.1109/TIP.2014.2303636

Ou, Y.-F., Zhou, Y., & Wang, Y. (2010). Perceptual quality of video with frame rate variation: A subjective study. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 2446–2449). IEEE. https://doi.org/10.1109/ICASSP.2010.5496300

Pedersen, M., Bonnier, N., Hardeberg, J., & Albregtsen, F. (2010). Attributes of image quality for color prints. *Journal of Electronic Imaging*, *19*(1), 011016. https://doi.org/10.1117/1.3277145

Picard, D., Dacremont, C., Valentin, D., & Giboreau, A. (2003). Perceptual dimensions of tactile textures. *Acta Psychologica*, *114*(2), 165–184. https://doi.org/10.1016/j.actpsy.2003.08.001

Pinson, M. H., & Wolf, S. (2003). Comparing subjective video quality testing methodologies. In T. Ebrahimi & T. Sikora (Eds.), *Visual Communications and Image Processing 2003* (Vol. 5150, pp. 573–582). Lugano, Switzerland. https://doi.org/10.1117/12.509908

Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., … Kuo, C.-C. J. (2014). Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, (0). https://doi.org/http://dx.doi.org/10.1016/j.image.2014.10.009

Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Astola, J., Carli, M., & Battisti, F. (2009). TID2008-A database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern …*, *10*, 30–45. Retrieved from http://www.ponomarenko.info/tid2008.htm

Radun, J., Leisti, T., Häkkinen, J., Ojanen, H., Olives, J.-L., Vuori, T., & Nyman, G. (2008). Content and Quality: Interpretation-based Estimation of Image Quality. *ACM Transactions on Applied Perception*, *4*(4), 1–15. https://doi.org/10.1145/1278760.1278762

Radun, J., Leisti, T., Virtanen, T., Häkkinen, J., Vuori, T., & Nyman, G. (2010). Evaluating the multivariate visual quality performance of image-processing components. *ACM Transactions on Applied Perception*, *7*(3), 1–16. https://doi.org/10.1145/1773965.1773967

Radun, J., Leisti, T., Virtanen, T., Nyman, G., & Häkkinen, J. (2014). Why is quality estimation judgment fast? Comparison of gaze control strategies in quality and difference estimation tasks. *Journal of Electronic Imaging*, *23*(6), 061103. https://doi.org/10.1117/1.JEI.23.6.061103

Radun, J., Nuutinen, M., Leisti, T., & Häkkinen, J. (2016). Individual differences in image-quality estimations: Estimation rules and viewing strategies. *ACM Transactions on Applied Perception*, *13*(3), 1–22. https://doi.org/10.1145/2890504

Radun, J., Virtanen, T., Nyman, G., & Olives, J.-L. (2006). Explaining multivariate image quality - Interpretation-Based Quality Approach. In *Proc. ICIS '06* (pp. 119–121). Rochester, New York, USA.

Radun, J., Virtanen, T., Olives, J.-L., Vaahteranoksa, M., Vuori, T., & Nyman, G. (2007). Audiovisual quality estimation of mobile phone video cameras with interpretation-based quality approach. In *Proc. SPIE 6494, Image Quality and System Performance IV* (Vol. 6494). San Jose, CA, United States: ociety of Photo-Optical Instrumentation Engineers (SPIE). https://doi.org/10.1117/12.703317

Ramanath, R., Snyder, W. E., Yoo, Y., & Drew, M. S. (2005). Color image processing pipeline. *IEEE Signal Processing Magazine*, *22*(1), 34–43. https://doi.org/10.1109/MSP.2005.1407713

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing Fluency and Aesthetic Pleasure: Is Beauty in the Perceiver's Processing Experience? *Personality and Social Psychology Review*, *8*(4), 364–382. https://doi.org/10.1207/s15327957pspr0804_3

Redi, J., Liu, H., Zunino, R., & Heynderickx, I. (2011). Interactions of visual attention and quality perception. In B. E. Rogowitz & T. N. Pappas (Eds.), *Human Vision and Electronic Imaging XVI. Proc. of SPIE-IS&T Electronic Imaging* (Vol. 7865, p. 78650S). San Francisco, (CA). https://doi.org/10.1117/12.876712

Redi, J., Zhu, Y., de Ridder, H., & Heynderickx, I. (2015). How Passive Image Viewers Became Active Multimedia Users. In C. Deng, L. Ma, W. Lin, & K. N. Ngan (Eds.), *Visual Signal Quality Assessment* (pp. 31–72). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-10368-6_2

Rehman, A., & Wang, Z. (2012). Reduced-Reference Image Quality Assessment by Structural Similarity Estimation. *IEEE Transactions on Image Processing*, *21*(8), 3378–3389. https://doi.org/10.1109/TIP.2012.2197011

Rossi, P. E., Gilula, Z., & Allenby, G. M. (2001). Overcoming scale usage heterogeneity: A bayesian hierarchical approach. *Journal of the American Statistical Association*, *96*(453), 20–31. https://doi.org/10.1198/016214501750332668

Rummukainen, O., Radun, J., Virtanen, T., & Pulkki, V. (2014). Categorization of Natural Dynamic Audiovisual Scenes. *PLoS ONE*, *9*(5), e95848. https://doi.org/10.1371/journal.pone.0095848

Saad, M. A., Bovik, A. C., & Charrier, C. (2010). A DCT Statistics-Based Blind Image Quality Index. *IEEE Signal Processing Letters*, *17*(6), 583–586. https://doi.org/10.1109/LSP.2010.2045550

Saad, M. A., Bovik, A. C., & Charrier, C. (2014). Blind Prediction of Natural Video Quality. *IEEE Transactions on Image Processing*, *23*(3), 1352–1365. https://doi.org/10.1109/TIP.2014.2299154

Säämänen, T., Virtanen, T., & Nyman, G. (2010). Videospace: classification of video through shooting context information. In S. P. Farnand & F. Gaykema (Eds.), *Proceedings of SPIE - The International Society for Optical Engineering* (Vol. 7529, p. 752906). https://doi.org/10.1117/12.839414

Segur, R. (2000). Using photographic space to improve the evaluation of consumer cameras. In *(PICS) Image Processing, Image Quality, Image Capture, Systems* (pp. 221–224). Portland, OR, United States: IS&T The Society for Imaging Science and Technology.

Seshadrinathan, K., Soundararajan, R., Bovik, A. C., & Cormack, L. K. (2010). Study of Subjective and Objective Quality Assessment of Video. *IEEE Transactions on Image Processing*, *19*(6), 1427–1441. https://doi.org/10.1109/TIP.2010.2042111

Sheikh, H. R., & Bovik, A. C. (2004). Image information and visual quality. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 3, pp. iii-709–712). IEEE. https://doi.org/10.1109/ICASSP.2004.1326643

Sheikh, H. R., Sabir, M. F., & Bovik, A. C. (2006). A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. *IEEE Transactions on Image Processing*, *15*(11), 3440–3451. https://doi.org/10.1109/TIP.2006.881959

Shibata, T., Kurihara, S., Kawai, T., Takahashi, T., Shimizu, T., Kawada, R., ... Nyman, G. (2009). Evaluation of stereoscopic image quality for mobile devices using interpretation based quality methodology. In A. J. Woods, N. S. Holliman, & J. O. Merritt (Eds.), *Stereoscopic Displays and Applications XX* (Vol. 7237, p. 72371E). https://doi.org/10.1117/12.807080

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, *74*(11), 1–29. https://doi.org/10.1037/h0093759

Staelens, N., Van Wallendael, G., Van de Walle, R., De Turck, F., & Demeester, P. (2013). High definition H.264/AVC subjective video database for evaluating the influence of slice losses on quality perception. In *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)* (pp. 130–135). IEEE. https://doi.org/10.1109/QoMEX.2013.6603225

Streijl, R. C., Winkler, S., & Hands, D. S. (2016). Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, *22*(2), 213–227. https://doi.org/10.1007/s00530-014-0446-1

Temel, D., & AlRegib, G. (2015). A comparative study of quality and content-based spatial pooling strategies in image quality assessment. In *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)* (pp. 732–736). Orlando, FL, USA: IEEE. https://doi.org/10.1109/GlobalSIP.2015.7418293

Tervonen, A., Nivala, I., Ryytty, P., Saari, H., Ojanen, H., & Viinikanoja, J. (2006). Integrated measurement system for miniature camera modules. In A. Tervonen, M. Kujawinska, W. IJzerman, & H. De Smet (Eds.), *SPIE 6196 Photonics in Multimedia, Photonics Europe* (Vol. 6196, p. 61960L). Strasbourg, France: Society of Photo-Optical Instrumentation Engineers (SPIE). https://doi.org/10.1117/12.662650

Teunissen, K. (1996). The Validity of CCIR Quality Indicators Along a Graphical Scale. *SMPTE Journal*, *105*(3), 144–149. https://doi.org/10.5594/J04650

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286. https://doi.org/10.1037/h0070288

Tinio, P. P. L., Leder, H., & Strasser, M. (2011). Image quality and the aesthetic judgment of photographs: Contrast, sharpness, and grain teased apart and put together. *Psychology of Aesthetics, Creativity, and the Arts*, *5*(2), 165–176. https://doi.org/10.1037/a0019542

To, M., Lovell, P. ., Troscianko, T., & Tolhurst, D. . (2008). Summation of perceptual cues in natural visual scenes. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1649), 2299–2308. https://doi.org/10.1098/rspb.2008.0692

Torgerson, W. S. (1958). *Theory and methods of scaling*. (1st ed.). New York, NY, USA: Wiley.

van Dijk, A. M., Martens, J.-B., & Watson, A. B. (1995). Quality asessment of coded images using numerical category scaling. In N. Ohta, H. U. Lemke, & J. C. Lehureau (Eds.), *Proc. SPIE* (Vol. 2451, pp. 90–101). https://doi.org/10.1117/12.201231

Virtanen, T., Nuutinen, M., & Häkkinen, J. (2019). Image quality wheel. *Journal of Electronic Imaging*, *28*(1), 1–12. https://doi.org/10.1117/1.JEI.28.1.013015

Virtanen, T., Nuutinen, M., Radun, J., Leisti, T., & Häkkinen, J. (2015). Alternative performance metrics and target values for the CID2013 database. In M.-C. Larabi & S. Triantaphillidou (Eds.), *Proceedings of SPIE*

- *The International Society for Optical Engineering* (Vol. 9396, p. 93960Q). https://doi.org/10.1117/12.2079100

Virtanen, T., Nuutinen, M., Vaahteranoksa, M., Oittinen, P., & Häkkinen, J. (2015). CID2013: A Database for Evaluating No-Reference Image Quality Assessment Algorithms. *IEEE Transactions on Image Processing*, *24*(1), 390–402. https://doi.org/10.1109/TIP.2014.2378061

Virtanen, T., Radun, J., Lindroos, P., Suomi, S., Säämänen, T., Vuori, T., … Nyman, G. (2008). Forming valid scales for subjective video quality measurement based on a hybrid qualitative/quantitative methodology. In S. P. Farnand & F. Gaykema (Eds.), *SPIE 6808, Image Quality and System Performance V* (Vol. 6808, p. 68080M). San Jose, CA, United States: Society of Photo-Optical Instrumentation Engineers (SPIE). https://doi.org/10.1117/12.765831

VQEG. (2000). VQEG FR-TV Phase I Database. Retrieved October 3, 2019, from https://www.its.bldrdoc.gov/vqeg/projects/frtv-phase-i/frtv-phase-i.aspx

VQEG. (2010). VQEG HDTV Database. Retrieved October 3, 2019, from https://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx

Vranješ, M., Rimac-Drlje, S., & Grgić, K. (2013). Review of objective video quality metrics and performance comparison using different databases. *Signal Processing: Image Communication*, *28*(1), 1–19. https://doi.org/10.1016/j.image.2012.10.003

Vu, C., Phan, T., & Chandler, D. (2012). S3: A Spectral and Spatial Measure of Local Perceived Sharpness in Natural Images. *IEEE Transactions on Image Processing*, *21*(3), 934–945. https://doi.org/10.1109/TIP.2011.2169974

Vu, P., & Chandler, D. (2012). A Fast Wavelet-Based Algorithm for Global and Local Image Sharpness Estimation. *IEEE Signal Processing Letters*, *19*(7), 423–426. https://doi.org/10.1109/LSP.2012.2199980

Vu, P., & Chandler, D. (2014). ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *Journal of Electronic Imaging*, *23*(1), 013016. https://doi.org/10.1117/1.JEI.23.1.013016

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. https://doi.org/10.1109/TIP.2003.819861

Winkler, S. (2005). *Digital Video Quality: Vision Models and Metrics*. Wiley.

Winkler, S. (2009). On the properties of subjective ratings in video quality experiments. In *2009 International Workshop on Quality of Multimedia Experience* (pp. 139–144). San Diego, CA: IEEE. https://doi.org/10.1109/QOMEX.2009.5246961

Winkler, S., & Dufaux, F. (2003). <title>Video quality evaluation for mobile streaming applications</title>. In T. Ebrahimi & T. Sikora (Eds.), *Proceedings of SPIE Visual Communications and Image Processing* (Vol. 5150, pp. 593–603). Lugano, Switzerland. https://doi.org/10.1117/12.509910

Wueller, D. (2006). *Image Engineering digital camera tests*. *Image Engineering* (Vol. 49).

Wueller, D., Artmann, U., Rao, V., Reif, G., Kramer, J., & Knauf, F. (2018). VCX: An industry initiative to create an objective camera module evaluation for mobile devices. *Electronic Imaging*, *2018*(5), 172-1-172–175. https://doi.org/10.2352/ISSN.2470-1173.2018.05.PMII-172

Wueller, D., Matsui, A., & Katoh, N. (2019). Visual Noise Revision for ISO 15739. *Electronic Imaging*, *2019*(10), 315-1-315–317. https://doi.org/10.2352/ISSN.2470-1173.2019.10.IQSP-315

Xu, J., Ye, P., Liu, Y., & Doermann, D. (2014). No-reference video quality assessment via feature learning. In *2014 IEEE International Conference on Image Processing (ICIP)* (pp. 491–495). IEEE. https://doi.org/10.1109/ICIP.2014.7025098

Yang, D., Peltoketo, V.-T., & Kämäräinen, J.-K. (2019). CNN-based Cross-dataset No-reference Image Quality Assessment. In *IEEE International Conference on Computer Vision (ICCV)*.

Yang, X., Li, F., & Liu, H. (2019). A Survey of DNN Methods for Blind Image Quality Assessment. *IEEE Access*, *7*, 123788–123806. https://doi.org/10.1109/ACCESS.2019.2938900

Yendrikhovskij, S., Blommaert, F., & de Ridder, H. (1999). Color reproduction and the naturalness constraint. *Color Research & Application*, *24*(1), 52–67. https://doi.org/10.1002/(SICI)1520-6378(199902)24:1<52::AID-COL10>3.0.CO;2-4

Yendrikhovskij, S., de Ridder, H., Fedorovskaya, E., & Blommaert, F. (1997). Colourfulness judgments of natural scenes. *Acta Psychologica*, *97*(1), 79–94. https://doi.org/10.1016/S0001-6918(97)00025-5

Yendrikhovskij, S., MacDonald, L., Bech, S., & Jensen, K. (1999). Enhancing colour image quality in television displays. *The Imaging Science Journal*, *47*(4), 197–211. https://doi.org/10.1080/13682199.1999.11736360

Zarzo, M., & Stanton, D. T. (2009). Understanding the underlying dimensions in perfumers' odor perception space as a basis for developing meaningful odor maps. *Attention, Perception & Psychophysics*, *71*(2), 225–247. https://doi.org/10.3758/APP.71.2.225

Zhang, F., Li, S., Ma, L., Wong, Y., & Ngan, K. (2011). IVP subjective quality video database. Retrieved October 3, 2019, from http://ivp.ee.cuhk.edu.hk/research/database/subjective

Zhang, L., Zhang, L., Mou, X., & Zhang, D. (2011). FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Transactions on Image Processing*, *20*(8), 2378–2386. https://doi.org/10.1109/TIP.2011.2109730

Zhang, Y., & Chandler, D. (2013). No-reference image quality assessment based on log-derivative statistics of natural scenes. *Journal of Electronic Imaging*, *22*(4), 043025. https://doi.org/10.1117/1.JEI.22.4.043025

Zhou, J., & Glotzbach, J. (2007). Image Pipeline Tuning for Digital Cameras. In *2007 IEEE International Symposium on Consumer Electronics* (pp. 1–4). IEEE. https://doi.org/10.1109/ISCE.2007.4382143

Zhu, J., & Wang, N. (2012). Image Quality Assessment by Visual Gradient Similarity. *IEEE Transactions on Image Processing*, *21*(3), 919–933. https://doi.org/10.1109/TIP.2011.2169971

ISBN 978-951-51-6361-5
UNIGRAFIA
HELSINKI 2020